

# Learning contextualized semantics from co-occurring terms via a Siamese architecture

Ubai Sandouk, Ke Chen\*

School of Computer Science, University of Manchester, Manchester, M13 9PL, UK



## ARTICLE INFO

### Article history:

Received 11 August 2015  
Received in revised form 28 October 2015  
Accepted 13 January 2016  
Available online 22 January 2016

### Keywords:

Contextualized semantics  
Descriptive terms  
Siamese architecture  
Out of vocabulary  
Semantic priming  
Representation learning

## ABSTRACT

One of the biggest challenges in Multimedia information retrieval and understanding is to bridge the semantic gap by properly modeling concept semantics in context. The presence of out of vocabulary (OOV) concepts exacerbates this difficulty. To address the semantic gap issues, we formulate a problem on learning contextualized semantics from descriptive terms and propose a novel Siamese architecture to model the contextualized semantics from descriptive terms. By means of pattern aggregation and probabilistic topic models, our Siamese architecture captures contextualized semantics from the co-occurring descriptive terms via unsupervised learning, which leads to a concept embedding space of the terms in context. Furthermore, the co-occurring OOV concepts can be easily represented in the learnt concept embedding space. The main properties of the concept embedding space are demonstrated via visualization. Using various settings in semantic priming, we have carried out a thorough evaluation by comparing our approach to a number of state-of-the-art methods on six annotation corpora in different domains, i.e., MagTag5K, CAL500 and Million Song Dataset in the music domain as well as Corel5K, LabelMe and SUNDatabase in the image domain. Experimental results on semantic priming suggest that our approach outperforms those state-of-the-art methods considerably in various aspects.

© 2016 Elsevier Ltd. All rights reserved.

## 1. Introduction

Multimedia information retrieval (MMIR) is a collective terminology referring to a number of tasks involving indexing, comparison and retrieval of multimedia objects (Jaimes, Christel, Gilles, Sarukkai, & Ma, 2005). As media content is created at an exponential rate, it has become increasingly difficult to manage even personal repositories of multimedia so as to make MMIR more and more demanding. Moreover, users expect certain levels of MMIR services from web service providers such as YouTube and Flickr. In addition, information processing tasks in fields such as medicine (Müller, Michoux, Bandon, & Geissbuhler, 2004) and education (Chang, Eleftheriadis, & McClintock, 1998) benefit enormously from advances in MMIR. In general, the most challenging problem in MMIR is the so-called *semantic gap* (Smeulders, Worring, Santini, Gupta, & Jain, 2000), which stems from the difficulty in linking low-level media representation, e.g., computationally extractable features, to high-level semantic concepts describing

the media content, e.g., human-like understanding. Bridging this gap has motivated a number of approaches including feature extraction (Lew, Sebe, Djeraba, & Jain, 2006), user-inclusive design (Schedl, Flexer, & Urbano, 2013), and high-level context modeling (Marques, Barenholtz, & Charvillat, 2011). By modeling concepts, the use of *semantics*, i.e., the representation of high-level concepts and their interactions, leads to improvements in MMIR applications as well as the interpretability of the retrieved results (Kaminskas & Ricci, 2012). As a result, semantics acquisition and representation are critical in bridging the semantic gap. The richness, meaningfulness and applicability of semantics rely primarily on the sources of concept-level relatedness information. Examples of such sources include manually constructed knowledge graphs or ontologies (Kim, Scerri, Breslin, Decker, & Kim, 2008), automatically analyzed media content (Torralba, 2003) or well-explored collections of crowd-sourced descriptive terms or tags (Miotto & Lanckriet, 2012).

As one of the information sources, *descriptive terms*, including keywords, labels and other textual descriptions of media, have also been used in capturing the term-based semantics underlying co-occurring descriptive terms. Such semantics provides direct concept-level knowledge regarding the concerned multimedia objects. Typical applications include music crowd tagging services

\* Corresponding author.

E-mail addresses: [ubai.sandouk@cs.manchester.ac.uk](mailto:ubai.sandouk@cs.manchester.ac.uk) (U. Sandouk), [chen@cs.manchester.ac.uk](mailto:chen@cs.manchester.ac.uk) (K. Chen).

<http://dx.doi.org/10.1016/j.neunet.2016.01.004>

0893-6080/© 2016 Elsevier Ltd. All rights reserved.

### Nomenclature

Symbol	Definition
$x[i]$	The $i$ th element of vector $\mathbf{x}$
$ X $	The cardinality of the set $X$
$\Upsilon$	Document-term relatedness matrix
$\tau$	A single descriptive term
$\Gamma$	The collection of training descriptive terms
$\delta$	A single document consisting of $m$ descriptive terms
$\Delta$	The collection of training documents
$\phi$	A single topic produced from LDA analysis
$\Phi$	The set of topics produced from LDA analysis of the training dataset
$\mathbf{t}(\tau)$	A representation of the descriptive term $\tau$
$\mathbf{l}(\tau \delta)$	A representation of the local context of term $\tau$
$\mathbf{BoW}(\delta)$	The binary Bag of Words representation of document $\delta$
$\overline{\mathbf{BoW}}(\delta)$	The binary complement of $\mathbf{BoW}(\delta)$
$\mathbf{x}(\tau, \delta)$	The collective representation of a term in context, i.e. $(\mathbf{t}(\tau), \mathbf{l}(\tau \delta))$
$h$	Index of layer in the neural network
$W_h, \mathbf{b}_h$	Weight matrix and biases pertaining to layer $h$
$\mathbf{CE}(\tau \delta)$	Contextualized embedding representation of a term in context
$\mathbf{E}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$	Euclidean distance between two terms' CE representations
$\mathbb{E}$	The abbreviated notation of CE Euclidean distance
$\mathfrak{d}$	The abbreviated notation of KL-divergence
$\mathbb{S}$	KL-divergence based similarity metric for local context
$\tau_{\text{OoV}}$	Out of vocabulary (OOV) descriptive term
$\mathbf{t}(\tau_{\text{OoV}})$	OOV descriptive term representation
$\delta_{\text{iv}}$	In-vocabulary terms in a document containing an OOV term
$\mathbf{CE}(\mathbf{x}(\tau_{\text{OoV}}, \delta_{\text{iv}}))$	Feature-based semantic representation of an OOV term
$\mathbf{CE}(\tau_{\text{OoV}} \delta_{\text{iv}})$	Concept-based semantic representation of an OOV term

(Law, Settles, & Mitchell, 2010) and multi-object image dataset analysis (Rabinovich, Vedaldi, Galleguillos, Wiewiora, & Belongie, 2007). Thanks to crowd-sourced annotation (Turnbull, Barrington, & Lanckriet, 2008) and game-based tags collection (Law, Ahn, Dannenberg, & Crawford, 2007), large collections of descriptive terms are now available. Those term collections can be analyzed for occurring patterns to reveal concept-level relatedness and similarity. Term-based semantics is expected to be transferable since it is acquired from high-level concepts independent of any specific MMIR tasks. It is worth stating that term-based semantics is different from linguistic semantics. First of all, descriptive terms are not only words but also symbols, abbreviations and complete sentences, e.g., “r’n'b” (musical style), “90s” (musical type), “stack of books” (visual concept), and so on. Next, descriptive terms may have a domain specific meaning different from their common linguistic meaning, e.g., “rock” is genre in music (not an earth substance) and “horn” is an instrument in music but is also a visual concept in images. Finally, the vocabulary used for descriptive terms is subject to change in time and cannot be fixed to represent a closed set of concepts. Those distinctions limit the usability of available linguistic resources such as linguistic dictionaries and generic word embedding from capturing term-based semantics. Therefore, we believe that the rich semantics conveyed in descriptive terms should be better explored and exploited to bridge the semantic gap.

By close investigation of various descriptive terms collections, we observe that terms could be used differently to represent

various types of semantics and relatedness: (a) a term may have multiple meanings and the intended meaning cannot be decided unless the term co-occurs with other coherent terms, e.g., the term “guitar” can refer to an acoustic guitar when it co-occurs with terms like “strings”, “classical”, and so on, or to an electric guitar when it co-occurs with terms such as “metal”, “rock”, and so on; (b) different terms may intend the exact same meaning regardless of context, e.g., “drums” and “drumset”; (c) different terms may have either similar or different meaning depending on context, e.g., “trees” and “forest” convey similar concepts and have similar meaning in context of natural scene (conveying a concept of many trees) but “tree” is by no means similar to “forest” when used in description of an urban scene; (d) different terms may share partial meaning but have different connotations, e.g., “house” and “building” convey some similar concepts but “building” has a wider connotation; and (e) co-occurring terms may not have their meanings in singularity or in pair but in group only, e.g., {“wing”, “tail”, “metallic”} together define a concept of an airplane while {“leg”, “cat”, “tail”, etc.} collectively present a concept of a cat and its body parts. The observations described above indicate the complexity and the necessity of taking the context into account in semantic learning from terms. Obviously, simply counting co-occurrence (Rabinovich et al., 2007) is insufficient in modeling various types of semantics and relatedness in descriptive terms to capture accurate concepts, and more sophisticated techniques are required so that we can capture all the intended semantics, or concepts and their relatedness, in descriptive terms accurately.

In general, a set of  $m$  terms,  $\delta = \{\tau_i\}_{i=1}^m$ , is often used collectively to describe the semantics underlying a single multimedia object where  $\tau_i$  is a descriptive term and  $\delta$  is the collective notation of the  $m$  terms, named *document* hereinafter. Furthermore, all  $m$  terms appearing in a document  $\delta$  are dubbed as *accompany* terms. Our observation reveals that for a specific term  $\tau_i$  in a document  $\delta$ , the accompany terms jointly create its contextual niche, named *local context*, that helps inferring the accurate intended meaning of  $\tau_i$  in that situation. In other words, the term along with its local context uniquely defines a concept of the accurate meaning. By taking such local contexts into account, we would learn a new type of relatedness between terms, named *contextualized relatedness*, by exploring terms' co-occurrence in different documents in a collection. Unlike the *global relatedness* where relatedness of terms is fixed irrespective of their local contexts, the contextualized relatedness of two terms is subject to change in the presence of different local contexts. In order to represent such contextualized semantics, we would embed all terms in a concept representation space that reflects the contextualized relatedness of terms. Formally, this emerging problem is formulated as follows: given a term  $\tau$  and its accompany terms in  $\delta$ , we would establish a mapping:  $(\mathbf{t}(\tau), \mathbf{l}(\tau|\delta)) \rightarrow \mathbf{CE}(\tau|\delta)$ , where  $\mathbf{t}(\tau)$  and  $\mathbf{l}(\tau|\delta)$  are the feature vectors of the term  $\tau$  and its local context in  $\delta$  and  $\mathbf{CE}(\tau|\delta)$  is a concept embedding representation of  $\tau$  given its local context in  $\delta$ , so that the contextualized semantic similarity of terms be properly reflected via a distance metric in the concept embedding representation space. This is a challenging problem due to the actual facts as follows: (a) terms get their meaning in groups rather than in singularity or in pair; (b) it is unclear how to capture intrinsic context in terms; and (c) terms that are not seen in training may appear in application runtime and hence may confuse a semantic learning model, this issue is known as *out-of-vocabulary* (OOV) issue in literature. Nevertheless, solving this problem brings us closer to bridging the semantic gap as a solution to this problem not only provides a term-level contextualized semantic representation, named *concept embedding* (CE) hereinafter, for a term to grasp an accurate concept as well as contextualized concept relatedness but also the representations of co-occurring terms in a document collectively form a novel document-level representation precisely modeling the concepts in groups

as well as subtle differences among those coherent concepts. Furthermore, the CE representation learnt from descriptive terms would facilitate a number of non-trivial applications including different MMIR tasks, e.g., auto-annotation of multimedia objects by mapping from the low-level visual/acoustic features onto the CE space, semantic retrieval by using the embedding representations as indexing terms, generating useful recommendations on both term and document levels in a recommendation system, and zero-shot learning in different multimedia classification tasks, e.g., object recognition and music genre classification.

In order to tackle the problem described above, we propose a novel Siamese architecture (Bromley et al., 1993) and a two-stage learning algorithm to capture contextualized semantics from descriptive terms. The proposed Siamese architecture learns the contextualized semantic embedding in an unsupervised way. The resultant CE representation space embeds different descriptive terms so that their contextualized semantic relatedness is reflected by their Euclidean distances. In this CE representation space, one term tends to co-locate with all the accompany terms appearing in its local context or co-occurring terms in the same document. As a result, our approach leads to multiple representations for a single term that appears in various documents, which reflects the polysemous aspects of a tag in different contexts. Thanks to our contextualized semantic learning, it becomes possible that the CE representation of an OOV tag can be inferred or approximated by using its local context, which paves a new way in solving the well-known OOV problem in MMIR. The semantics learnt in this way is also naturally generic yet transferable as it does not rely on any specific MMIR tasks. Depending on the nature of descriptive terms used in practice, the semantics acquired from some training collections may also be domain specific. With different training and test corpora, we would verify the above-mentioned transferability and domain-specific properties of our CE representation generated in our experiments of various settings.

Our main contributions in this paper are summarized as follows: (a) we formulate a problem for learning contextualized semantics from co-occurring descriptive terms and propose a novel Siamese architecture and a two-stage learning algorithm as a solution to this problem; (b) we propose two treatments based on our CE representation to address the issues regarding OOV terms; (c) we demonstrate the main properties of our CE representation via visualization; and (d) by means of semantic priming, we thoroughly evaluate the performance of our CE representation with a number of various settings by comparing to several state-of-the-art semantic learning methods.

The rest of the paper is organized as follows. Section 2 reviews the related work in terms of learning different relatedness from descriptive terms. Section 3 describes feature extraction of term and local context required in our contextualized semantic learning. Section 4 presents our Siamese architecture and learning algorithms. Section 5 describes the experiments on the CE learning with our Siamese architecture and Section 6 presents experimental settings and results in semantic priming. Section 7 discusses relevant issues and Section 8 draws conclusions.

## 2. Related work

In this section, we review relevant works in learning semantics from descriptive terms regardless of any specific multimedia tasks. In terms of semantics learnt from descriptive terms, those approaches fall into one of three different categories: *global*, *syntactic* and *contextualized* relatedness.

### 2.1. Global relatedness

Global relatedness refers to the relatedness between pairs of terms that does not take any context into account. In general, there are *statistics-based* and *graph-based* methodologies for learning global relatedness from descriptive terms.

Aggregation (Markines et al., 2009) is a statistical-based method that focuses on pairwise co-occurrence of terms in the training dataset and is sometimes named co-occurrence analysis. By considering all training documents, aggregation works on a document-term matrix  $\mathcal{Y}$  where the presence or absence of each term in each document is represented as binary or frequency indicator (Singhal, 2001). Thus, a column of document-term matrix  $\mathcal{Y}$  forms a feature vector for the use of one term. As relatedness between pairs of terms is likely reflected in their pair-wise use pattern, it can be estimated by measuring the distance between the corresponding terms' feature vectors. Hence, the relatedness between pairs of terms can be learnt from a training set with statistical measures. As the relatedness is obtained from an entire dataset, it is not affected by local context. As an extension, the term-to-term relatedness matrix achieved with all the pair-wise relatedness may be further analyzed with *Principle Component Analysis* (PCA). By removing unwanted redundancy and noise, the resultant term representation is in a lower dimensional space. This extension yields improved performance in the movie review sentiment evaluation task (Lebret, Legrand, & Collobert, 2013). Nevertheless, such extension is sensitive to preprocessing and tunable parameters. Similarly, Mandel et al. (2011) proposed an *information theoretic* inspired (InfoTheo) method that yields a smoothed document representation. InfoTheo directly alters the values of  $\mathcal{Y}$  in favor of terms that co-occur frequently across an entire dataset. This smoothed representation is later aggregated in order to generate a term-to-term relatedness matrix. Nevertheless, this smoothing process introduces more parameters and hence results in heavier parameter tuning.

Another statistic-based method is *Latent Semantic Indexing* (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSI is often used to analyze collections of documents with large vocabulary or descriptive terms. In LSI, the matrix  $\mathcal{Y}$  is decomposed using *Singular Value Decomposition* (SVD) as  $\mathcal{Y}^T = U\Sigma V^T$ . Two orthogonal matrices  $U$  and  $V$  correspond to the terms and the document subspaces, respectively. The dimensionality of the subspaces is controlled by limiting the entries of the diagonal matrix  $\Sigma$ , i.e., retaining the first few entries only. This decomposition generates an approximation of  $\mathcal{Y}$  with the smallest reconstruction error. It also uncovers collective terms' usage patterns. The rows of  $U$  can readily be used as representations of the descriptive terms. The relatedness between the terms is measured by the cosine similarity between their corresponding vectors (Levy & Sandler, 2008). Unfortunately, LSI generally suffers from poor generalization to new terms/documents.

Graph-based models rely on a graph representation where terms are mapped to nodes and pair-wise relatedness is mapped to an edge between relevant nodes. Such a graph may be constructed by manual or automatic relatedness analysis (Huetting, Monszpart, & Mellado, 2014; Kim et al., 2008; Wang, Anguera, Chen, & Yang, 2010). However, graph-based models are subject to capacity limitation; any additional node representing a new term has to be introduced manually in graph revision. Moreover, an edge representing relatedness usually has a fixed cost that does not take the local context into account. Therefore, graph-based models are often thought of as handcrafted dictionaries of semantics.

In summary, those approaches to learning global relatedness do not address the issue of the contextualized semantics but yield a term-level representation efficiently. In our work, we would explore such approaches in generating a semantic representation of terms required by our learning model (see the next section for details).



## 2.2. Syntactic relatedness

In natural languages, context is explicitly present in the order of the words, i.e., syntactic context. This dependency between sequences of words helps capturing the words' relatedness in context and understanding of basic linguistic meanings consequently.

To capture the syntactic relatedness, distributed language models (Collobert et al., 2011; Mikolov, Corrado, Chen, & Dean, 2013; Mikolov, Karafiát, Burget, Cernocký, & Khudanpur, 2010) have been proposed recently. Such models learn syntactic relatedness from linguistic corpora and yield distributed semantics where words are embedded based on their syntactic similarity (Mikolov et al., 2010). During learning, a model is trained to predict a missing word given some context, e.g., nearby words, or to predict possible context words given a word. If trained properly, interchangeable words without breaking language rules, i.e., syntactically close words, would have close embedding vectors. Those models have attracted increasing attention due to their simplicity and capacity in providing generic semantics for various tasks (Frome et al., 2013; Mikolov et al., 2013). Moreover, Pennington, Socher, and Manning (2014) showed how to combine the advantages of PCA models with this syntactic relatedness by careful analysis of the ratios of co-occurrence probabilities between pairs of words appearing in each other's neighborhood.

Although language models yield a contextualized representation, they entirely rely on the syntactic context and hence are not applicable to descriptive terms where there is no synthetic dependency and the co-occurring terms may describe a multimedia object regardless of their orders. Nevertheless, such techniques can be employed as a baseline in a thorough evaluation of the contextualized semantics learned from descriptive terms studied in this paper.

## 2.3. Contextualized relatedness

Motivated by syntactic relatedness, terms are permitted to exhibit varying inter-relatedness levels depending on the context. Works in this stream focus on document-level representations where patterns of terms' use are captured in a *document-level* representation. Consequently, measuring similarity between documents may be straightforward while terms' meaning and their relatedness are difficult to capture. This often hinders the applicability of such models as generic semantics providers. Here, we review approaches that can potentially capture contextual relatedness studied but all lead to only a document-level representation.

Topic models are a class of statistical methods used for semantics modeling. A topic model makes use of latent processes to capture collective occurrence patterns in the form of statistical distributions over observed terms, called *topics*. When a specific term appears in more than one document, which exhibits different patterns of use with other terms, it might be associated with more than one topic, which suggests its different meanings stochastically. Those multiple term-topic associations capture the different levels of relatedness among terms.

*Latent Dirichlet Allocation* (LDA) (Blei, Ng, & Jordan, 2003) and *Probabilistic Latent Semantic Analysis* (PLSA) (Hofmann, 1999) are the most prominent topic models used in text and natural language processing. In LDA and PLSA, a set of independent topics  $\Phi$  are used to softly cluster the documents based on the used terms. During learning, the process estimates scalar priors  $B$  for the Dirichlet distribution in LDA or Multinomial distribution in PLSA that models the topics as distributions over terms as well as the scalar prior  $B^0$  that models the topics distribution. After training, the posterior probability of all topics given a term and the topic probability given a document can be estimated with the trained models. Given a

term  $\tau$ , the posterior probability of a topic  $\phi_c \in \Phi$  is  $p(\phi_c|\tau) \sim p(\tau|\phi_c)p(\phi_c)$ ; where  $p(\tau|\phi_c) \sim \text{Categorical}(\text{Dirichlet}(B_c))$  and  $p(\phi_c) \sim \text{Dirichlet}(B^0)$  in LDA or  $p(\tau|\phi_c) \sim \text{Categorical}(B_c)$  and  $p(\phi_c) \sim \text{Uniform}(B^0)$  in PLSA. Given a document  $\delta$ , the topic probability is  $p(\phi_c|\delta) \sim p(\phi_c) \prod_{\tau \in \delta} p(\tau|\phi_c)$ .

While both LDA and PLSA capture contextualized semantics, such models provide a summary of a document in form of a mixture of topics; they capture ad hoc relatedness but do not provide term-to-term relatedness explicitly. As a result, the relatedness between a pair of terms  $\tau_1$  and  $\tau_2$  has to be estimated under a specific topic distribution:  $\theta(\delta) = \{p(\theta_c) = p(\phi_c|\delta)\}_{c=1}^{|\Phi|}$ . Assuming equal priors for all the terms, the relatedness between two terms,  $\tau_1$  and  $\tau_2$ , might be estimated by the Kullback–Leibler (KL) divergence:

$$\text{KL}(\tau_1, \tau_2|\theta(\delta)) = \sum_{c=1}^{|\Phi|} \frac{p(\theta_c)}{p(\tau_1)} (p(\tau_1|\theta_c) - p(\tau_2|\theta_c)) \cdot \left( \log \left( \frac{p(\tau_1|\theta_c)}{p(\tau_2|\theta_c)} \right) \right).$$

Thus, the term-topic relatedness learnt by LDA or PLSA implicitly contextualizes the relatedness between terms under different topic distributions. Without considering term-to-term relatedness directly, however, semantics yielded by LDA or PLSA is encoded in a collective term representation rather than a concept-level representation required by a solution to our problem. The effectiveness of topic models in learning semantics from descriptive terms has been evaluated in Law et al. (2010) and Levy and Sandler (2008). They reported increased accuracy over global relatedness models when performing auto tagging of music. In their settings, the task does not require a term-to-term relatedness measure and hence a topic-to-term relatedness offered by topic models is sufficient.

Another method for introducing a context is using *Conditional Restricted Boltzmann Machines* (CRBM) in Mandel et al. (2011). CRBM (Taylor, Hinton, & Roweis, 2007) is a variant of the traditional RBM (Hinton, 2002) consisting of visible layer  $\mathbf{v}$  and hidden layer  $\mathbf{h}$ . The probabilistic units in visible and hidden layers are fully connected via a weight matrix  $U$ , and vectors  $\mathbf{d}$  and  $\mathbf{c}$  are the biases in visible and hidden layers, respectively. For learning in RBM, an energy function is defined by:

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{h}^T U \mathbf{v} - \mathbf{d}^T \mathbf{v} - \mathbf{c}^T \mathbf{h}.$$

The model is trained by minimizing the free energy  $F(\mathbf{v}) = -\log \sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})}$  with the contrastive divergence (CD) algorithm.

In CRBM (Mandel et al., 2011), an additional visible layer for condition  $\mathbf{a}$  is introduced and connected to the original visible layer via a weight matrix  $W$ . As a result, the energy function in CRBM is defined

$$E(\mathbf{v}, \mathbf{h}, \mathbf{a}) = -\mathbf{h}^T U \mathbf{v} - \mathbf{v}^T W \mathbf{a} - \mathbf{d}^T \mathbf{v} - \mathbf{c}^T \mathbf{h}.$$

CRBM can also be trained by the CD algorithm. When it is used in learning semantics from descriptive terms, its observed vector  $\mathbf{v}$  is set to be the *Bag of Words* (BoW) binary representation (Harris, 1954) of a concerned document. One binomial unit is used for each vocabulary term. The hidden vector  $\mathbf{h}$  is set to binomial variables that capture occurrence patterns. The condition  $\mathbf{a}$  is set to the one-hot representation of the training documents where exactly one unit is used as index representation of a document. The collection of terms used by other users for the same document can also be used as another condition in Mandel et al. (2011) if available. During test, the term-to-term relatedness is measured by co-activation between a query term and other terms. In this process, the unit corresponding to one query term is clamped "on" (as well as the relevant condition unit) and sampling chains of a large number of times are undertaken.

Eventually, the average activation level of each visible unit encodes relatedness of its corresponding term to the query term under the context conditions. While CRBM provides a smoothed relevance of each term to a relevant document, the semantics captured is limited to a document-level representation rather than concept-level relatedness studied in this paper. In particular, this approach suffers from a fundamental weakness as it does not lead to a deterministic continuous semantic embedding representation required by various applications.

In summary, the existing works for learning semantics from descriptive terms focus on only relatedness of specific types and does not sufficiently address the issues arising from our formulated problem. To verify our argument, we have used all the approaches reviewed above as baselines in our semantic priming experiments (see Section 6 for details).

### 3. Term, local context and document representation

In this section, we describe feature extraction for the descriptive term, local context and document representations employed in our approach to facilitate the presentation of our proposed Siamese architecture in the next section.

#### 3.1. Term representation

In general, terms can be characterized by either *ID-based* or *statistics-based* representations. An ID-based representation uses a scheme directly linked to the term's ID in symbolic form, i.e., a separate entity for each term. A statistics-based representation uses statistical analysis of the used terms within the dataset. The ID-based term representations have been used in previous models, e.g., the CRBM model (Mandel et al., 2011). However, the capacity of an ID-based representation may be limited so that adding new terms becomes difficult. Moreover, it may impose an unnatural order on terms, e.g., a numerical ID. Therefore, we employ a statistics-based representation (Markines et al., 2009) where each descriptive term is represented as a summary of its pair-wise use with all terms over an entire training dataset. This summary encodes the global relatedness among pairs of terms and can work together with the local context to form a raw concept representation of a term in context as described below.

To achieve the statistics-based representation, we start from the training document-term matrix with binary entries described in Section 2.1. In our work, we do not eliminate any terms in a training dataset as we believe that the entire collection of terms in documents form a coherent meaning niche conveying proper local contexts collectively. The document-term binary matrix is re-weighted using *tfidf* which highlights those rarely used terms. Given a vocabulary of descriptive terms  $\Gamma$  and a training dataset  $\Delta$ , the binary *term frequency* of the presence of the term  $\tau \in \Gamma$  in a document  $\delta \in \Delta$  is found in the corresponding entry in the document-term matrix:

$$tf(\tau, \delta) = \begin{cases} 1 & \text{when } \tau \text{ appears in } \delta \\ 0 & \text{otherwise.} \end{cases}$$

The rarity of a term  $\tau$  in the collection is achieved by the inverted document frequency *idf* ( $\tau$ ):

$$idf(\tau) = \log \left( \frac{|\Delta|}{1 + |\{\delta : \tau \text{ appears in } \delta\}|} \right),$$

where  $|\cdot|$  is the cardinality of a set.

Then *tfidf* ( $\tau, \delta$ ) = *tf* ( $\tau, \delta$ )  $\times$  *idf* ( $\tau$ ).

After reweighting the matrix, each term is described using all *tfidf* values of its use and is represented by its usage vector  $\mathbf{u}(\tau) =$

$\{tfidf(\tau, \delta_i)\}_{i=1}^{|\Delta|}$ . Then, the global relatedness between two terms  $\tau_1$  and  $\tau_2$  is obtained by aggregation with the dot product:

$$T(\tau_1, \tau_2) = \langle \mathbf{u}(\tau_1), \mathbf{u}(\tau_2) \rangle.$$

Thus, a term is represented by a feature vector of  $|\Gamma|$  features consisting of its global relatedness to all terms in the training dataset:

$$\mathbf{t}(\tau) = \{T(\tau, \tau_i)\}_{i=1}^{|\Gamma|}. \quad (1)$$

#### 3.2. Local context representation

As described in Section 1, the local context of each term is acquired by considering all the terms in the same document as they together convey specific concepts. A local context representation should be semantically consistent and easy to capture in real applications, e.g., auto annotation. In the recent work of Law et al. (2010), Latent Dirichlet Allocation (LDA) was used to represent terms in form of topics and then a model was trained to map the acoustic content onto the topical representation to facilitate MMIR. Motivated by their work, we employ LDA to represent the local context in our work due to the generality of LDA in representing patterns of collective use. It is worth mentioning that there are alternative models for local context representations, e.g., semantic hierarchies, PLSA or any other topic model. Here, we emphasize that a local context representation used in our work is not equivalent to the complete document itself but a semantically coherent summary of the document.

To achieve the local context representation with LDA (see Section 2.3 for more details), a set of topics  $\Phi$  softly clusters the documents based on the used terms within each document. During training, the process estimates scalar priors  $B$  for Dirichlet distributions modeling the topics as distributions over terms as well as scalar prior  $B^0$  modeling the topic distribution itself. After training, the probability of observing a term  $\tau \in \Gamma$  given a specific topic  $\phi \in \Phi$  follows

$$p(\tau|\phi) \sim \text{Categorical}(\text{Dirichlet}(B)),$$

where  $p(\phi) \sim \text{Dirichlet}(B^0)$ .

This means that the probability of one term identifying one topic follows  $p(\phi|\tau) \sim p(\tau|\phi)p(\phi)$  and the likelihood of a topic given a complete document  $\delta$  follows  $p(\phi|\delta) \sim p(\phi) \prod_{\tau \in \delta} p(\tau|\phi)$ . Consequently, given a term and accompany terms, the local context is represented by a feature vector of  $|\Phi|$  features corresponding to  $|\Phi|$  topic distribution output:

$$\mathbf{l}(\tau|\delta) = \{l_c(\delta)\}_{c=1}^{|\Phi|}, \quad l_c(\delta) = p(\phi_c|\delta). \quad (2)$$

#### 3.3. Document representation

Apart from term and local context representations, a representation of an entire document is also required in our approach. In our work, we adopt the *Bag of Words* (BoW) representation of a document  $\delta$  denoted by  $\mathbf{BoW}(\delta)$ , a binary sparse feature vector of  $|\Gamma|$  entries for a given vocabulary  $\Gamma$ , where entry  $i$  corresponds to a specific term  $\tau_i$ :

$$\mathbf{BoW}(\delta)[i] = \begin{cases} 1 & \text{when } \tau_i \text{ appears in } \delta \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

In summary, we employ the *tfidf*-based aggregation as our term representation to encode the global term-to-term relatedness, the LDA as our local context representation to summarize semantic coherence in different documents and the BoW as the document representation in our learning model.

#### 4. Model description

In this section, we come up with a solution to the problem described in Section 1. We first describe our motivation behind our proposed Siamese architecture and then present its architecture and a two-stage algorithm for learning contextualized semantics from descriptive terms. Finally, we propose two methods to deal with the contextualized semantic embedding of OOV terms based on the representation space generated by our Siamese architecture.

##### 4.1. Motivation

As described in Section 1, we aim to tackle an issue that has not been fully addressed previously in learning semantics from descriptive terms. In the previous work, either the term-to-term relatedness is captured without taking the local context into account or the context is modeled on a document level. Unlike previous work, we encounter a challenge where a term and its local context have to be simultaneously taken into account. By looking into the nature of this problem, we would like to come up with a solution by fulfilling two subsequent tasks.

In general, an ideal representation of semantics allows similar concepts to associate each other seamlessly; a concept should be easily inferred from its related/coherent concepts. Motivated by the argument that learning a simple yet relevant auxiliary task could facilitate semantic embedding learning (Bottou, 2014), we can comply with this requirement by fulfilling a simple yet generic task: *predicting all the accompany terms in a document from the representations of a constitutional term and its local context* described in Section 3. If a learning model of latent variables is employed, we expect that the latent variables form a representational space that encodes the semantic information of coherent terms at a concept level. As such a representation also needs to resolve the highly nonlinear relationship between terms and their contexts in order to predict accompany terms, a deep neural network of hidden layers would be a powerful tool for this task.

While the representation generated by the prediction task encodes the semantic information conveyed in coherent terms, it may not provide the proper term-to-term relatedness in context. To enhance the semantic representational space, we need to perform a further task based on the initial semantic representation obtained in the prediction task; i.e., *learning a proper distance metric for the pairwise contextualized relatedness of concepts*. For this task, we would develop a variant of Siamese architecture consisting of two identical deep neural networks used in the earlier prediction task as a Siamese architecture has turned out to be an effective method for distance learning (Bromley et al., 1993). By taking all possible concept relations between a pair of terms along with their local contexts into account during learning, we expect that all concepts reflected by terms in the presence of local contexts are located properly in the embedding space so that a pair of coherent concepts sharing the local context can co-locate with minimal distance and other concepts can be positioned properly in reflection of their contextualized relatedness.

Upon accomplishing the proposed learning tasks, we anticipate that all contextualized concepts are properly embedded in a distributed representation space.

##### 4.2. Architecture

As illustrated in Fig. 1, the proposed Siamese architecture consists of two identical subnetworks. Each subnetwork is a feed-forward neural network composed of  $H - 1$  hidden layers and two visible layers marked in green; i.e., input and output layers.

Each subnetwork receives the representations of a term  $\mathbf{t}(\tau)$  and its local context  $\mathbf{l}(\tau|\delta)$ , collectively denoted by  $\mathbf{x}(\tau, \delta) = (\mathbf{t}(\tau), \mathbf{l}(\tau|\delta))$ , as input and outputs a prediction of the BoW representation of all the terms in  $\delta$ ,  $\mathbf{BoW}(\delta)$ , in the document  $\delta$  from which  $\mathbf{t}(\tau)$  and  $\mathbf{l}(\tau|\delta)$  were extracted. Two subnetworks are coupled to work together and trained via a two-stage learning procedure.

In the first stage, one subnetwork is trained to carry out the prediction task for an initial semantic embedding. As a result, this subnetwork is trained to predict the BoW representation of a document,  $\mathbf{BoW}(\delta)$  from the input features of a tag  $\tau$  and its local context in  $\delta$ ,  $\mathbf{x}(\tau, \delta)$ . After the first-stage learning, the output of the  $(H - 1)$ th hidden layer is used as an initial contextualized semantic representation for concepts conveyed by terms and their local contexts. We refer to this representation as concept embedding (**CE**) throughout the paper.

In the second stage, we couple two identical trained subnetworks and train two subnetworks simultaneously to revise the initial semantic embedding towards embedding the proper contextualized term-to-term or concept-to-concept relatedness. The learning in this stage is done via *distance learning* working on further constraints required by a proper distance metric in contextualized semantic embedding. During the distance learning, two subnetworks work together to deal with different situations regarding all possible types of input to two subnetworks. For regularization, each subnetwork is also trained simultaneously in this stage to perform the prediction task in order to avoid unnecessary changes to initial semantic representation achieved in the first stage with a multi-objective optimization process.

After the two-stage learning, we achieve two identical subnetworks. Those are used in mapping a term and its context to the CE space to form its contextualized representation.

##### 4.3. Learning algorithm

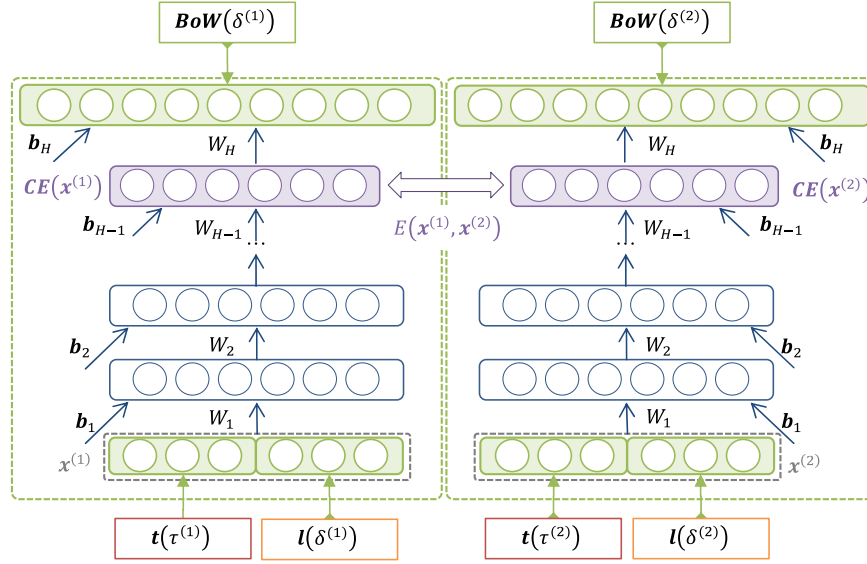
To facilitate the presentation of our learning algorithm, we first describe our notation system (see also Nomenclature). For layer number  $h$  in a subnetwork, the output is

$$\mathbf{z}_h(\mathbf{x}) = \mathbf{f}(W_h \cdot \mathbf{z}_{h-1}(\mathbf{x}) + \mathbf{b}_h), \quad 1 \leq h \leq H,$$

where  $W_h, \mathbf{b}_h$  are the weights and bias vectors for the  $h$ th layer of the network,  $\mathbf{f}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$  is the element-wise hyperbolic tangent function. We stipulate that  $\mathbf{z}_0(\mathbf{x}) = \mathbf{x}$  indicates the input layer,  $\mathbf{CE}(\mathbf{x}) = \mathbf{z}_{H-1}(\mathbf{x})$  is the contextualized semantic representation vector, i.e., the output of the  $(H - 1)$ th hidden layer, and  $\hat{\mathbf{y}}(\mathbf{x}) = \mathbf{z}_H(\mathbf{x})$  is the prediction vector yielded by the output layer. Hereinafter, we shall drop all the explicit parameters to simplify the presentation, e.g.,  $\mathbf{y}_k$  is an abbreviated version of  $\mathbf{y}_k(\mathbf{x}_k(\tau, \delta))$  and  $y_k[j]$  denotes the  $j$ th entry of vector  $\mathbf{y}_k$ .

##### 4.3.1. Training data

For unsupervised learning, we need to create training examples based on different documents in a term collection or collections used for training. Given a training document  $\delta$  consisting of  $m$  co-occurring terms, we create  $m$  training examples where each example is a focused term in document with the same local context, i.e., the  $m$  terms in the document  $\delta$ . The prediction targets for all the  $m$  examples are the same, i.e. the document representation of this training document  $\mathbf{BoW}(\delta)$ . We observed that in training for the prediction, the local context may predominate the initial semantic embedding and hence cause all terms in the same document to have very similar representations regardless of whether they are meaningfully coherent. To tackle this issue, we introduce *negative* examples. Given a training document  $\delta$ , we synthesize a negative example by randomly



**Fig. 1.** The proposed Siamese architecture for learning contextualized semantics from descriptive terms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

coupling a term that is not in  $\delta$  and using all the terms in  $\delta$  to form its local context. The prediction target for such a negative example is set to be the complement of a document representation of  $\delta$ ; i.e., the complement of  $\text{BoW}(\delta)$  denoted by  $\overline{\text{BoW}(\delta)}$ , achieved by flipping all the binary entries of  $\text{BoW}(\delta)$ . To avoid confusion, hereinafter, we refer to those examples with prediction target  $\text{BoW}(\delta)$  as *positive* examples. For a balanced learning, we use all positive examples and the same number of randomly synthesized negative examples. Thus, for any example  $k$ , its input is  $\mathbf{x}_k(\tau, \delta) = (\mathbf{t}(\tau), \mathbf{l}(\tau|\delta))$  and the learning target is the document representation of  $\delta$ , i.e.,  $\mathbf{y}_k(\mathbf{x}_k(\tau, \delta)) = \text{BoW}(\delta)$  if  $\mathbf{x}_k(\tau, \delta)$  is a positive example or  $\mathbf{y}_k(\mathbf{x}_k(\tau, \delta)) = \overline{\text{BoW}(\delta)}$  otherwise.

#### 4.3.2. Prediction learning

To learn the prediction in the first stage, a *deep neural network* (DNN) is initialized with the greedy unsupervised layer-wise pre-training procedure using sparse auto-encoders as building blocks (Bengio, Lamblin, Popovici, & Larochelle, 2007). After a subnetwork of  $H - 1$  hidden layers is initialized, the DNN is fine-tuned by applying the document representation labels. The learning algorithm for sparse autoencoder can be found in the Appendix.

The binary nature of the output makes the cross-entropy loss suitable for this task. Given the entire training dataset  $(X, Y)$  of  $K$  examples generated from  $|\Delta|$  documents and a vocabulary of  $|\Gamma|$  terms, the initial prediction loss is

$$\mathcal{L}_P(X, Y; \Theta) = -\frac{1}{2K|\Gamma|} \sum_{k=1}^K \sum_{j=1}^{|\Gamma|} ((1 + y_k[j]) \log(1 + \hat{y}_k[j]) + (1 - y_k[j]) \log(1 - \hat{y}_k[j])),$$

where  $\Theta$  is the collective notation of all the weight and bias parameters in the DNN. This is a standard loss function for binary classification and the targets in our situation are binary BoW representations. During learning, however, the sparse nature of the BoW representation often skews the target labels towards an incorrect trend that all the terms are absent in a document. To escape from this trend, we re-weight the cost of a false negative error for example  $k$ , i.e., the existing term is predicted as absence, by  $\kappa_k = |\{j : y_k[j] = 1\}| / |\Gamma|$ . By re-weighting errors caused by different examples, the loss incurred by false negative errors

in learning is highlighted. Thus, we use the following re-weighted prediction loss in prediction learning:

$$\mathcal{L}_P(X, Y; \Theta) = \frac{-1}{2K|\Gamma|} \sum_{k=1}^K \sum_{j=1}^{|\Gamma|} (\kappa_k (1 + y_k[j]) \log(1 + \hat{y}_k[j]) + (1 - \kappa_k) (1 - y_k[j]) \log(1 - \hat{y}_k[j])). \quad (4)$$

Solving this optimization problem based on a training dataset leads to a trained deep neural network that can predict all the accompany terms in a document from their term and local context representations. Details of this learning algorithm can be found in the Appendix.

#### 4.3.3. Distance learning

After completing the prediction learning with a single DNN, we train a Siamese architecture by coupling two copies of the trained DNNs. When presenting a pair of input vectors  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$  of two examples to the coupled DNNs, we employ Euclidean distance between their CE representations (cf. Fig. 1) to measure the embedding similarity:

$$E(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \|\mathbf{CE}(\mathbf{x}^{(1)}) - \mathbf{CE}(\mathbf{x}^{(2)})\|_2. \quad (5)$$

Although other distance metric can be used for the same purpose, we wish to use the simplest Euclidean distance to be our learning goal. Furthermore, local context intrinsically decides the semantic meaning of different concepts. Thus, measuring the similarity between two different local contexts is essential for our distance learning. As described in Section 3.2, the local context similarity in the LDA can be measured by the Kullback–Leibler (KL) divergence:

$$\text{KL}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sum_{c=1}^{|\Phi|} \left( (l^{(1)}[c] - l^{(2)}[c]) \log \left( \frac{l^{(1)}[c]}{l^{(2)}[c]} \right) \right). \quad (6)$$

To formulate a loss for our distance learning, we exploit the information underlying all the paired training examples used to train the Siamese architecture. According to our observations, two terms (in the same document) sharing the same local context often convey the same concept. Hence, the Euclidean distance between their representations in the CE space should be zero ideally. On the other hand, for two terms (in different documents)



of different local contexts, their Euclidean distance between the representations in the CE space should be set via learning to reflect their local context similarity measured by the KL divergence. Hence, we have to enumerate all different situations that incur the loss in our distance learning.

Let the binary variables  $I_1$ ,  $I_2$  and  $I_3$  indicate three possible situations for a pair of examples that have input  $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ :

- $I_1 = 1$ : both  $\mathbf{x}^{(1)}$  and  $\mathbf{x}^{(2)}$  are positive examples. In this situation, the proper distance between their representations is learnt to reflect the conceptual similarity between their local contexts. In particular, their representations of two terms sharing the same local context should be co-located or as close as possible in the CE space.
- $I_2 = 1$ : both  $\mathbf{x}^{(1)}$  and  $\mathbf{x}^{(2)}$  are negative examples. In this situation, the same should be done as described for  $I_1 = 1$ . As the concepts conveyed in negative examples are randomly synthesized, however, the distance between their CE representations is less important. This difference will be reflected by a treatment of weighting the similarity cost differently for  $I_1 = 1$  and  $I_2 = 1$  in our loss as presented below.
- $I_3 = 1$ :  $\mathbf{x}^{(1)}$  and  $\mathbf{x}^{(2)}$ ; one is positive and the other is negative. In this situation, the distance between their CE representations is unknown given the fact that the coherence is deeply uncertain for a positive example conveying certain concept and a negative example is randomly synthesized, which may not convey any concepts. Thus, their CE representations need to be distant as far as possible to distinguish from positive examples conveying the genuine concepts especially when they share the same local context. Again, we carry it out with a weighting scheme in our loss as presented below.

As a result, our loss for distance learning needs to take all the above three situations into account alternately with different types of pair-examples as three situations are mutually exclusive; i.e.,  $\sum_{i=1}^3 I_i = 1$  and  $I_i \in \{1, 0\}$  for  $i = 1, 2, 3$ .

Given two subsets  $X^{(1)}$  and  $X^{(2)}$  of the same cardinality  $N$  of examples from the training dataset via random pairing. For example pair  $n$ , we denote  $\mathbb{E} = E(\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(2)})$ ,  $\mathbb{d} = \text{KL}(\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(2)})$  and  $\mathbb{S} = e^{-\frac{\lambda}{2}\mathbb{d}}$  where  $\lambda$  is a positive sensitivity hyper-parameter. While  $\mathbb{E}$  and  $\mathbb{d}$  are defined in Eqs. (5) and (6),  $\mathbb{S}$  forms a local context similarity measure carried out by an exponential decay of the KL distance  $\mathbb{d}$  and the hyper-parameter  $\lambda$  decides the degree to which the embedding is dominated by the local context similarity, as required in dealing with three different situations described above. If two terms share the same local context, i.e.,  $\mathbb{d} = 0$ , the resultant similarity  $\mathbb{S} = 1$ . Otherwise the similarity reflects the actual difference between two local contexts. Hence, we define our loss based on three situations indicated by  $I_1$ ,  $I_2$  and  $I_3$  as follows:

$$\begin{aligned} \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta) = & \sum_{n=1}^N (I_1 (\mathbb{E} - \beta(1 - \mathbb{S}))^2 \\ & + I_2 \rho (\mathbb{E} - \beta(1 - \mathbb{S}))^2 \\ & + I_3 (\mathbb{E} - \beta)^2 \mathbb{S}). \end{aligned} \quad (7)$$

Here,  $\beta$  is a scaling hyper-parameter used to ensure controlled spreading concepts over the entire CE space and  $\rho$  is a hyper-parameter that weights down the importance of the situation of  $I_2 = 1$ . Intuitively, the first two terms in Eq. (7) define the precise loss between the actual distance  $\mathbb{E}$  and the target distance  $\beta(1 - \mathbb{S})$  for situations  $I_1 = 1$  and  $I_2 = 1$  but the importance of the loss for situation  $I_2 = 1$  is discounted by  $\rho$ . The last term in Eq. (7) specifies a loss that penalizes a negative example to be co-located with a positive example via a cost  $\mathbb{E} - \beta$  along with considering their local context similarity  $\mathbb{S}$ .

In the distance learning, we still need to keep the CE space of the main properties resulting from the prediction learning. As a result, the loss used in our distance learning is a multi-objective cost function by combining losses defined in Eqs. (4) and (7):

$$\begin{aligned} \mathcal{L}(X^{(1)}, X^{(2)}, Y^{(1)}, Y^{(2)}; \Theta) = & \sum_{i=1}^2 \mathcal{L}_p(X^{(i)}, Y^{(i)}; \Theta_i) \\ & + \alpha \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta), \end{aligned} \quad (8)$$

where  $\alpha$  is used as a trade-off hyper-parameter to balance two losses and  $\Theta_i$  is a collective notation of all the parameters in one of two component DNNs.

In our two-stage learning, all the parameters are estimated iteratively via the stochastic back-propagation (SBP) (Bottou, 2012) by optimizing the loss functions specified in Eqs. (4) and (8). The optimal hyper-parameters are found via a grid search with cross-validation and, in general, an early stopping criterion is applied in the SBP learning (see the next section for our specific experimental setting). In each iteration of the SBP, a small batch of training examples are randomly selected to update the parameters in training either single DNN for the prediction learning or Siamese architecture for the distance learning. It is also worth stating that the two component DNNs in Siamese architecture are always made identical via averaging their weights and biases after each iteration during the distance learning. Details of our learning algorithms and their derivation can be found in the Appendix.

#### 4.4. OOV term contextualized embedding

Upon applying the contextualized semantic representation learned from co-occurring terms, the issue of *out-of-vocabulary* (OOV) term contextualized embedding has to be addressed. In general, a test document may contain more than one OOV term. As our learning model relies on the LDA in generating the local context representation but the LDA does not address the OOV issue, we always use only those in-vocabulary terms appearing in this test document when generating the LDA-based local context representation. Without loss of generality, we need to take a document containing only one OOV term into account. Based on the trained Siamese architecture and the resultant contextualized semantic representation, we propose two methods to deal with the OOV situation. Let  $\delta = \{\tau_{oov}, \delta_{iv}\}$  denote a test document of an OOV term  $\tau_{oov}$  where  $\delta_{iv} = \{\tau_i\}_{i=1}^m$  is a collective notation of  $m$  in-vocabulary terms in  $\delta$ .

Our first method relies on the representation capacity of the term representation described in 3.1. We can extend the term representation to an OOV term  $\tau_{oov}$ . For  $\tau_{oov}$ , its *tfidf* values are measured the same as done for in-vocabulary terms with all the training documents plus all test documents containing it. However, the aggregation is only done against all the in-vocabulary terms to achieve  $\mathbf{t}(\tau_{oov})$ . This extension ensures that the same number of features is used to represent both in-vocabulary and OOV terms. In addition, each feature always means relatedness against the corresponding term. Similarly, we can achieve its local LDA-based context representation  $\mathbf{l}(\tau_{oov}|\delta_{iv})$  by using  $\delta_{iv}$ . Thus, feeding  $\mathbf{x}(\tau_{oov}, \delta_{iv}) = (\mathbf{t}(\tau_{oov}), \mathbf{l}(\tau_{oov}|\delta_{iv}))$  to a trained subnetwork leads to its contextualized semantic representation  $\mathbf{CE}(\mathbf{x}(\tau_{oov}, \delta_{iv}))$ . As OOV terms were not seen in training, a document of any OOV term  $\delta = \{\tau_{oov}, \delta_{iv}\}$  is actually equivalent to the settings of negative examples (cf. Section 4.3.1). In our distance learning, we have made all negative examples distant from positive examples in the CE space as far as possible. When measuring the contextualized term-to-term relatedness between an OOV term and other in-vocabulary terms, we stipulate that its most related in-vocabulary term is the one furthest in distance in the CE space. As this treatment relies



on the term representation, we name it the *feature-based* OOV method.

Our second method is motivated by the coherent nature of co-occurring terms in a document and the capacity of our contextualized semantic representation in encoding term-level and document-level semantics. As a result, the contextualized semantic representation  $\mathbf{CE}(\tau_{oov}|\delta_{iv})$  should be co-located with  $\{\mathbf{CE}(\mathbf{x}(\tau_i, \delta_{iv}))\}_{i=1}^m$  in the CE space and shares the group-level semantics in the CE space. Thus, we directly define the contextualized semantic representation of  $\tau_{oov}$ :

$$\mathbf{CE}(\tau_{oov}|\delta_{iv}) = \frac{1}{m} \sum_{i=1}^m \mathbf{CE}(\mathbf{x}(\tau_i, \delta_{iv})).$$

Intuitively, we treat the OOV term as missing data and then use the centroid of  $\mathbf{CE}$  representations of  $m$  co-occurring in-vocabulary terms to represent the concept conveyed in this OOV in terms of their shared local context. As this process does not involve the OOV term features, we deliberately use the notation  $\mathbf{CE}(\tau_{oov}|\delta_{iv})$  to distinguish from the OOV term representation  $\mathbf{CE}(\mathbf{x}(\tau_{oov}, \delta_{iv}))$  achieved by the first method. As this treatment is based on the  $\mathbf{CE}$  representations of those accompany in-vocabulary terms in the document containing OOV terms, we dub it the *concept-based* OOV method.

## 5. Experiments on concept embedding learning

In this section, we describe the experimental settings and visualize results regarding the use of our Siamese architecture to learn the concept embedding (CE) space from a number of corpora in different domains.

### 5.1. Datasets

In our experiments, we employ six publicly accessible corpora of multi-term documents from two domains: textually tagged music and multi-labeled images. Those datasets are often used as benchmarks for different information processing tasks including MMIR.

Three music tagged corpora are CAL500 (Turnbull, Barrington, Torres, & Lanckriet, 2007), MagTag5K (Marques, Domingues, Langlois, & Gouyon, 2011) and Million Song Dataset (MSD) (Bertinmahieux, Ellis, Whitman, & Lamere, 2011). It is observed that the three music tagging datasets exhibit different yet typical aspects of music tagging. In our experiments, we used only CAL500 and MagTag5K, respectively, as our training corpora to investigate the influence of different tagging styles in our CE learning and the MSD as a test dataset to examine the generalization of contextualized semantics learnt from a specific dataset via cross-corpora setting (cf. Sections 5.2 and 6.2.2).

Three multi-labeled image datasets used in our experiments are Corel5K (Duygulu, Barnard, Freitas, & Forsyth, 2002), LabelMe dataset (Russell, Torralba, Murphy, & Freeman, 2007) and SUN-Database benchmark (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). Unlike the music annotation case, we observe comparable statistics underlying the annotations in the three image corpora. In particular, there exists similar usage statistics including label-use frequency and document cardinality in both Corel5K and LabelMe. In our experiments, we used only Corel5K as the training corpora in our CE learning and LabelMe and SUNDatabase as test datasets in the cross-corpora setting (cf. Sections 5.2 and 6.2.2).

In summary, Table 1 describes the information on six datasets including domain, the number of documents in a dataset (#Doc.), the number of in-vocabulary terms (#In-Voc.), the averaging document length, the number of OOV terms (#OOV) reserved for simulation (see also Sections 5.2 and 6.2.4 for details) and the purpose in our experiments.

### 5.2. Experimental setting

We now describe the experimental settings in training our Siamese architecture on three corpora: CAL500, MagTag5K and Corel5K.

For feature extraction, we applied methods described in Section 3 to generate the term, the local context and the document representation from each document. We can achieve the feature vectors of any in-vocabulary term with Eq. (1) and generate the representation of OOV terms in a similar way as described in Section 4.4. In our experiments, the de-correlation of features with PCA and linear scaling of each feature was applied to the term and the local context representations in order to ensure that each feature is in the range  $(-1, +1)$ . By applying Eq. (2), the local context features of a document were obtained based on a trained LDA working on all accompany terms in the document. To train an LDA model, we use all the documents in a training dataset. The number of topics were empirically decided by using the hierarchical process as suggested in Teh, Jordan, Beal, and Blei (2006). As a result, we achieved three LDA models of 25, 19 and 20 topics trained on CAL500, MagTag5K and Corel5K, respectively. Each LDA model is applied to a relevant document to generate its local context representation. Note that a trained parametric LDA model is also used in generating the local context representations for those test documents in different settings. For training the Siamese architecture, the BoW representation of a training document is achieved with Eq. (3).

For model selection and performance evaluation in different settings, cross validation (CV) was used. In CV, a training corpus is randomly split into two subsets A and B with a ratio 2:1; A for training and B for validation and test. For CAL500, 40 documents in B were randomly chosen and reserved for validation during training and the rest of documents in this subset were used for test. For MagTag5K, we adopted a default setting suggested by Marques, Domingues et al. (2011) instead of the random split and 300 documents were randomly selected from subset B for validation while the rest of documents were reserved for test. The same setting as done for MagTag5K was applied to Corel5K. Furthermore, it should be clarified that we have exploited MagTag5K in simulating OOV situations. To do so, we randomly reserved 22 tags from the MagTag5K vocabulary. Thus, the number of in-vocabulary tags is down to 114. Accordingly, all the documents containing any of those 22 tags are removed before the CV split. Hence, the number of documents used in the aforementioned CV setting is 3826.

For the Siamese architecture, we randomly generate the same number of negative examples as that of positive examples in subset A by using the procedure described in Section 4.2 and append them to subset A in each CV trial. It is worth mentioning that the use of more negative than positive examples often leads to a degenerate solution that the uniform negative output in prediction is always reached regardless of any actual input. For the distance learning, as described in Section 4.3, training documents in subset A were randomly paired so that roughly equal number of paired examples was generated for two situations corresponding to  $l_i = 1$  ( $i = 1, 2$ ). Consequently, the number of examples for  $l_3 = 1$  doubles that number.

In the SBP learning, the “optimal” hyper-parameter values were found with a grid search during multiple CV trials and summarized as follows: (a) for the sparse autoencoder learning, the sparsity factor is 2, weight decay is 0.02 and a quasi-Newton algorithm was employed for training (see Appendix for details); (b) for the prediction learning, the learning rates were initially set to  $10^{-4}$ ,  $10^{-5}$  and  $10^{-5}$  for MagTag5K, CAL500 and Corel5K, respectively, and then decayed with a factor of 0.95 every 200 epochs. The re-weighting parameter  $\kappa$  in Eq. (5) is automatically obtained for each example as described in Section 4.3.2; and (c) for the distance

**Table 1**

Summary of all the datasets used in our experiments.

Dataset	Domain	#Doc.	#In-Voc.	Ave. Doc. length	#OOV	Training
CAL500	Music	500	158	25	0	✓
MagTag5K	Music	5,259	136	5	22	✓
MSD	Music	218,754	24,499	8.5	N/A	×
Corel5K	Images	4,524	292	3.5	0	✓
LabelMe	Images	26,945	2,385	7.3	N/A	×
SUNDatabase	Images	23,743	1,908	11	N/A	×

learning, the importance and the scaling factors in Eq. (6) were set to  $\rho = 0.5$  and  $\beta = \sqrt{d}$ , respectively, and the trade-off factors in Eq. (7) were set to  $\alpha = 2000$  for MagTag5K and  $\alpha = 1000$  for both CAL500 and Corel5K. The same learning rates used in the prediction learning were used for distance learning but the decay rule was applied every 200 mini-batches in SBP.

Early stopping principle was applied in both the prediction and the distance learning stages for generalization. Instead of monitoring only the cost defined in the loss functions on a validation set, however, our stopping criterion makes use of a surrogate loss on the validation set; i.e., the performance of a semantic priming task,  $P@2$ , to be described in the next section. The motivation behind this stopping criterion comes from the unsupervised learning nature of our CE learning; the loss functions were formulated for generic semantics without ground-truth. As a generic information retrieval task, semantic priming allows us to see “ground-truth” to some extent. Hence, the actual generalization performance can be guaranteed at least on the generic semantic priming task. As a result, our stopping criterion is as follows: we evaluated the priming performance based on the representations obtained after every 200 epochs and examined the performance improvement on both training and validation datasets between two adjacent tests. The learning was stopped at the point of the smallest improvement between two test points by human inspection. We believe that this is a generic stopping criterion applicable to any applications of our contextualized semantic representation. In Section 5.3, we demonstrate that this early stopping criterion actually leads to satisfactory concept embedding.

For model selection, we examined a number of feed-forward neural networks that have hidden layers ranging from one to four layers and different numbers of hidden units in a hidden layer ranging from 10 to 200. For reliability, we repeated the aforementioned CV experiments for three trials. As a result, the “optimal” subnetwork:  $input \rightarrow 100 \rightarrow 100 \rightarrow 10 \rightarrow output$ ; i.e., a multi-layered perceptron has three hidden layers of 100, 100 and 10 hidden units where the dimension of the CE representation is 10. It is worth mentioning that our model selection described above was mainly done based on the MagTag5K training set. For training on CAL500 and Corel5K, the grid search score is much smaller thanks to the information acquired from the MagTag5K training. Actually, the optimal structure achieved based on the MagTag5K training turns out to be the best for both CAL500 and Corel5K as well. Hereinafter, we report experimental results based on this optimal structure.

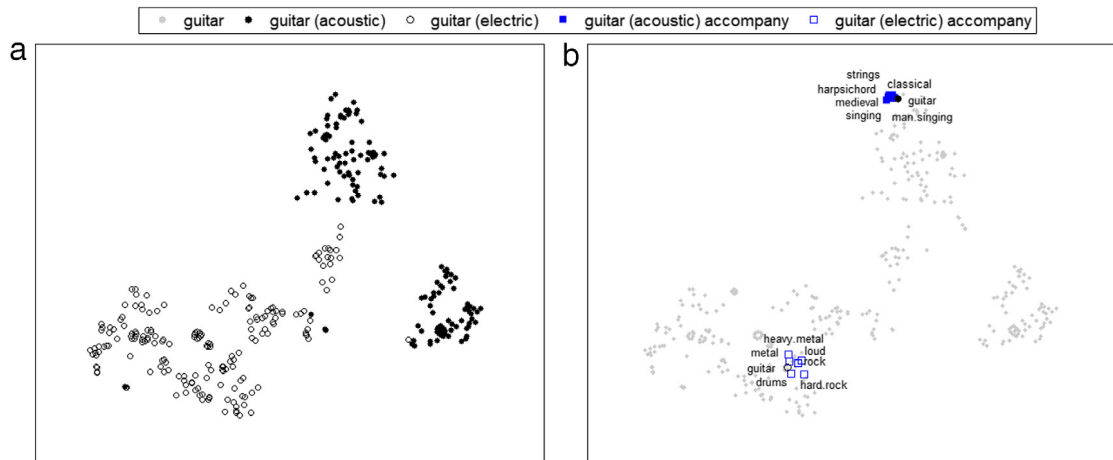
### 5.3. Visualization of concept embedding

After the completion of the two-stage learning, a trained sub-network provides a 10-dimensional CE representation for any given term along with its local context. By employing the unsupervised t-SNE (van der Maaten & Hinton, 2008), we can visualize the CE representations learnt from the training corpora by projecting the 10-dimensional representation to a 2-dimensional space. Thanks to the powerful non-linear dimensionality reduction capacity of the t-SNE, we anticipate that the visualization would

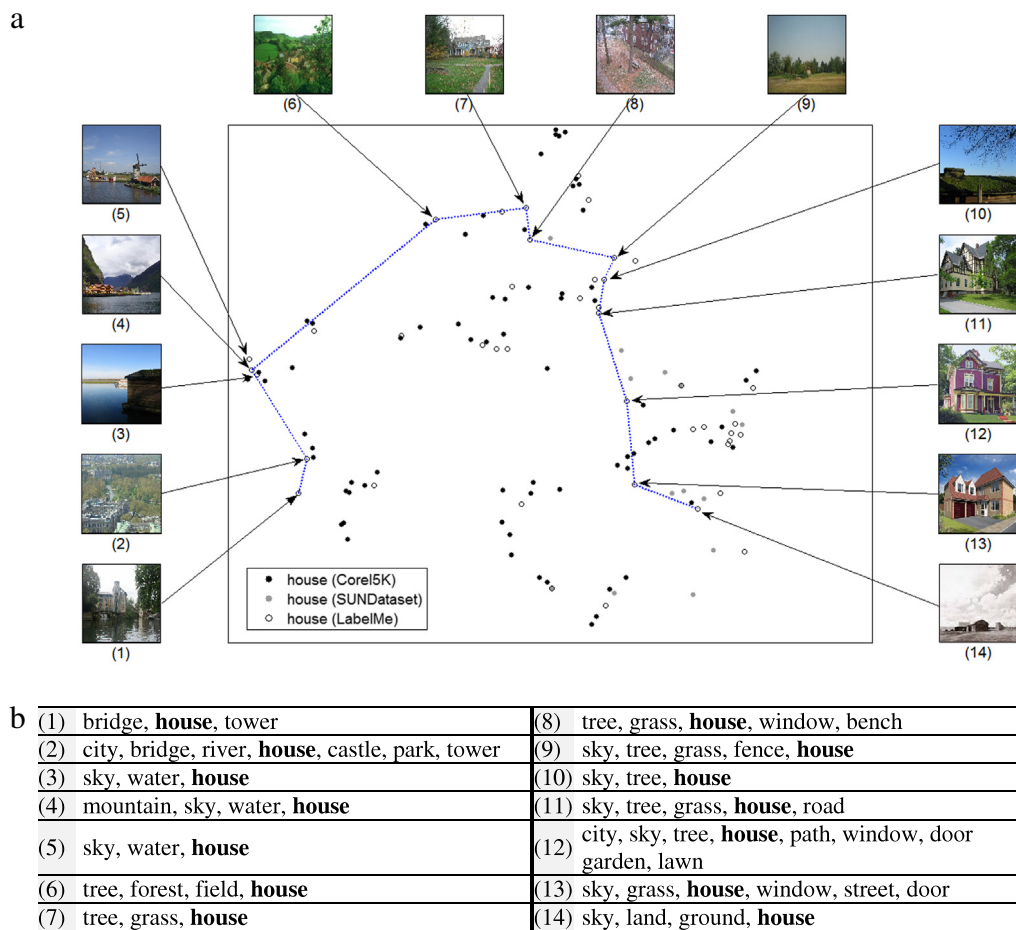
demonstrate the main properties of contextual semantics and relatedness learnt from training corpora in different domains vividly.

First of all, we choose “guitar” to be the focused tag as it is a typical example of a tag that can convey different concepts in the presence of different local contexts (cf. Section 1). We collect all the 388 documents containing “guitar” from the MagTag5K dataset and apply the subnetwork trained on MagTag5K to produce their CE representations for all 388 “guitar” with different local contexts. To facilitate our presentation, hereinafter, *instance* is used to describe an embedding vector of a focused term along with its local context. Fig. 2 shows the projection of all the CE representations of 388 “guitar” instances onto 2-D space as well as the projection of CE representations of few relevant tags that share the same local context with the focused tag. It is observed from Fig. 2(a) that the concepts defined by the tag “guitar” instances are grouped into three clusters, which demonstrates that our CE representation captures multiple meanings of “guitar” in different contexts. By a closer look, we find that three clusters actually correspond to two different meanings or concepts: “acoustic guitar” indicated by the solid circle (●) and “electric guitar” indicated by the hollow circle (○). As two different instruments are often used in different music genres, our CE representation has successfully distinguished between them by embedding them in different regions. Fig. 2(b) further shows the projection of CE representations corresponding to the accompany tags from two randomly chosen documents containing “guitar”, one indicated by solid square (■) from an “acoustic” cluster and the other indicated by hollow square (□) from the “electric” cluster, by superimposing them on the projection of “guitar” as shown in Fig. 2(a). Note that we deliberately shade all 388 “guitar” instance projections in Fig. 2(b) for clearer visualization. It is clearly seen from Fig. 2(b) that different concepts in the same context have been properly co-located with each other in the CE space thanks to the distance learning used in training the Siamese architecture. Furthermore, the CE representations of all the co-occurring tags in a document collectively provide a document-level representation reflecting a set of similar concepts and their subtle differences. While such concepts in context seem to be easily grasped by people, we emphasize that the embedding was acquired via unsupervised learning.

Next, we take the label “house” in the image domain as an example to examine whether concepts reflecting the ambient environment specified in its local contexts can be captured by our CE representation. Moreover, we would demonstrate the transferability of learnt contextualized semantics via visualization. As a result, we collect all the documents containing the label “house” in Corel5K, LabelMe and SUNDatabase datasets and use the Siamese architecture trained on Corel5K to generate the CE representations for 152 “house” instances. Fig. 3 illustrates the projections of all 152 “house” instances in a 2-D space where 96 instances indicated by solid circle (●) from Corel5K, 14 instances indicated by gray circle (●) from SUNDatabase, and 42 instances indicated by hollow circle (○) from LabelMe. Due to the unavailability of images in Corel5K, we have to examine the embedding by inspecting all the relevant documents manually. In general, our inspection shows that the contextualized semantics



**Fig. 2.** Visualization of the *CE* representations of “guitar” in MagTag5K. (a) 2-D projections corresponding to all the 388 “guitar” instances. (b) 2-D projections of the instances corresponding to the accompany tags co-occurring with “guitar” in two randomly selected documents containing “guitar”. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



**Fig. 3.** Visualization of the *CE* representations of label “house” in image datasets. (a) 2-D projections of all 152 “house” instances along with the associated images of “house” instances in LabelMe whose projections are roughly located on a manifold as indicated by connected points. (b) Annotation documents used as local contexts for those instances on the manifold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

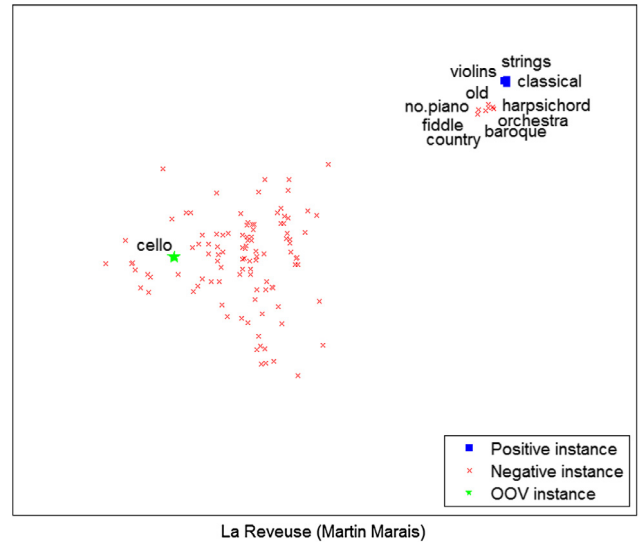
learnt from Corel5K properly reflects concepts corresponding to different ambient environments for the “house” instances in all three datasets and the 2-D projections of their *CE* representations are illustrated in Fig. 3. Fortunately, we can use images from LabelMe to confirm our inspection. As a result, we present 14 “house” images in Fig. 3(a) and the corresponding annotations in Fig. 3(b). It is evident that houses with similar ambient

environments are close to each other in the *CE* space. In particular, it is observed that a manifold appears in the 2-D space and shows the transition of ambient environments from castles, seaside and rural houses to urban houses. As illustrated in Fig. 3(a), we highlight the manifold by connecting those projection points on the “house” manifold in response to the ambient environmental changes. We highlight that the concepts are captured solely from

the co-occurring labels via unsupervised learning without using any visual features.

Finally, we demonstrate OOV term embedding via visualization. As described in Section 4.4, the concept-based OOV embedding method entirely relies on the representations of in-vocabulary terms appearing in the local context and the **CE** representation of an OOV term is actually the centroid of its co-occurring in-vocabulary term representations in the CE space. One easily imagines such an embedding. As a result, we simply visualize the **CE** representation of an OOV term achieved by the feature-based embedding method (cf. Section 4.4). In our experimental settings described in Section 5.2, 22 tags in MagTag5K were reserved to simulate OOV terms. For visualization, we choose a typical document of four tags {"classical", "violins", "strings", "cello"} that annotates the song "La Reveuse" composed by Martin Marais. In this document, "cello" is one of OOV terms. To facilitate our presentation of the main properties of an OOV term in the CE space, we also visualize all the instances derived from the incomplete document of {"classical", "violins", "strings"} including all the positive/negative examples. Consequently, the incomplete document leads to three positive and 111 negative instances by coupling all the remaining 111 in-vocabulary tags with this incomplete document (cf. Section 4.3.1). Hence, tags "classical", "violins" and "strings" are in turn to be the focused tags in three positive instances and the document containing this tags and {"classical", "violins" and "strings"} together form its local context. To generate a negative instance, we substitute the focused tag in the positive instance with an in-vocabulary tag other than "classical", "violins" and "strings". Fig. 4 illustrates 2-D projections of the **CE** representations of "cello" and all relevant instances specified above. It shows the projections of all the instances concerning the exemplar document as described above. It is observed from Fig. 4 that all three positive instances indicated by blue square (■) are co-located and projected onto a tiny region at the upper right corner of the 2-D space. With the music knowledge, we see that all three instances correspond to concepts that classical music is played by string instruments. In contrast, 111 negative instances indicated by cross in red (×) and are projected to two regions in the 2-D space; i.e., the small region consisting of seven instances is close to three positive instances and the large one composed of the remaining negative instances is far from the small region as well as those projections of three positive instances as shown in Fig. 4. A closer look at those near three positive instances reveals that the tags used to form those instances, as depicted in Fig. 4, are actually semantically associated with the positive instances (even though they are treated as negative). Any of those tags might have been used to annotate this music piece without altering the concept; i.e., most classical string-based music is orchestral in an old style, probably from the Baroque era and rarely involving piano in it. Moreover, most such music includes the fiddle as instrument. This result demonstrates the capability of our approach in capturing the accurate concepts underlying training documents even for those treated as "negative" in training. From Fig. 4, it is seen that the OOV tag instance ( $\tau_{oov} = \text{"cello"}$  and  $\delta_{iv} = \{\text{"classical", "violins", "strings"}\}$ ) indicated by green star (★) is projected into the large region of negative example due to the fact that OOV term was not seen in training and hence the resultant OOV instance has to be treated as negative (cf. Section 4.4). As the OOV instance is further from the three points corresponding to positive instances than any negative instances in the 2-D space, the visualization intuitively provides the evidence to support our measure of the contextualized relatedness between in-vocabulary and OOV terms in the feature-based OOV treatment.

In summary, visualization shown above suggests that our learning model successfully captures contextualized semantics from co-occurring terms in different domains and also demonstrates its



**Fig. 4.** 2-D projections of the **CE** representation of the OOV tag "cello" along with those sharing the same local context. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

capability in dealing with domain-specific semantics, transferability of learnt semantics across different corpora and the OOV terms. In addition, visualization also suggests that the use of a surrogate loss, i.e., semantic priming performance, in our stopping criterion during learning leads to generic **CE** representations applicable to various tasks described in Section 1.

## 6. Application to semantic priming

As demonstrated in Section 5, our trained model captures high quality terms semantics that tends to be generic and hence can support a variety of applications. *Semantic priming* is an application that depends directly on those semantics without the need of accessing content or other information regarding media (Lund & Burgess, 1996; Osgood, 1952). As priming highlights the versatility of the semantics from an abstract point of view, it provides invaluable insight into the performance of a semantic model regardless of different applications. Hence, we employ this generic task to evaluate the performance of our proposed approach based on those datasets described in Section 5.1 and further compare ours to a number of state-of-the-art methods on learning semantics from co-occurring terms for thorough evaluation.

### 6.1. Semantic priming and evaluation

In general, semantic priming is a process involving associating concepts based on their semantic relatedness. This abstract process is often used to evaluate the learnt semantics and demonstrate the performance of a semantic learning model (Lund & Burgess, 1996). Ideally, coherent terms should be associated with one another based on the intrinsic contextualized semantics conveyed by them. To do so, a semantic learning model has to resolve the highly nonlinear relationship between terms and contexts by capturing intentions behind those observed terms as accurately as possible. Thus, the semantic priming task becomes a proper test bed to evaluate the capabilities of a semantic learning model by measuring the relatedness of terms in different scenarios such as applicability to new documents, incomplete context and the presence of OOV terms.

Below, we first present the priming protocol used in evaluating a term-based contextualized semantic representation. Then, we extend this protocol to the document-level so that all the existing



semantic learning models can be compared fairly on the exact same condition. Finally, we describe the evaluation criteria used in semantic priming.

### 6.1.1. Priming protocol

Semantic priming was first introduced by Meyer and Schvaneveldt (1971) to associate semantically related concepts to each other, e.g., doctor–nurse. Through semantic priming, it has been shown that human subjects read consecutive words quicker when the words are semantically or syntactically associated. Given a priming concept or a query concept, all other concepts can be generally split into two groups, related and unrelated concepts, depending on the context. Thus, semantic priming functions as a highly generic evaluation method for learnt semantics without access to information other than the learnt semantics themselves. Semantic priming was first used in Lund and Burgess (1996) in evaluating the appropriateness of learnt *similarity* where a word embedding space was built using aggregation of a textual corpus. Using such embedding, word similarity was estimated using priming such that the closer two words' representations were, the more similar they were considered to be. However, there exists a subtle difference between syntactic and semantic relatedness (cf. Section 2.2) and we focus on the semantic relatedness in our experiments reported in this section. As a result, we define the priming as the capability of a semantic learning model in identifying related concepts given a single query concept represented as a term and its local context as defined in Section 1.

Priming is reflected by a learnt semantic representation where similarity between concepts is encoded in some semantic distance. In contextualized semantics, such distance is significantly affected by the context. A contextualized semantic model uses the context to express concepts meaningfully via their contextualized semantic distance denoted by  $e(\tau_1, \tau_2|\delta)$  where  $\tau_1$  and  $\tau_2$  are two different terms in the document  $\delta$  that forms their shared local context. This distance measure can straightforwardly be applied to priming over a set of concepts as follows: given a query concept  $\tau$ , all available terms  $\tau_i$  ( $i = 1, \dots, |\Gamma|$ ) in a vocabulary  $\Gamma$  are ranked based on their corresponding contextualized semantic distances to the query concept:

$$\text{Prime}(\tau, \delta) = (\tau_i | \forall \tau_i, \tau_j \in \Gamma : e(\tau, \tau_i|\delta) \leq e(\tau, \tau_j|\delta) \text{ if } i \leq j)_{i=1}^{|\Gamma|}. \quad (9)$$

Intuitively, Eq. (9) results in an ordered list of all  $|\Gamma|$  terms whose corresponding representations have increasing distances away from the query concept. Ideally, the terms with contextualized semantic similarity to the query concept should precede those sharing no such semantic similarity, and the top term of the list may correspond to the query concept itself.

For a semantic learning model, acquiring such ranked list for a specific query concept depends only on the definition of a distance metric used in its representation space. In literature, the priming performance evaluation requires ground-truth or gold standard of all the different concept similarities in terms of all possible contexts. Unfortunately, such information is not only missing for descriptive terms so far but also does not seem attainable in general since it demands the human judgment on terms' relatedness in an unlimited number of contexts.

To alleviate the problem of the ground-truth unavailability, we assume that all co-occurring terms in a single document are coherent and hence, semantically similar in terms of their shared context. Thus, one document is used as ground-truth; each time one of its constitutional terms is used as a query term to prime other terms in that document. As a result, the priming protocol used in our experiment is as follows. Given a document, each term in this document would be used in turn as a query term that

couples with the shared local context derived from the document to form a query concept. The list of primed terms resulting from each query term is then compared against this document (treated as ground-truth) to measure the priming accuracy as described in Section 6.1.3. The performance of a contextualized semantic learning model is evaluated by taking priming accuracy on all evaluation documents into account.

### 6.1.2. Extended priming protocol

The priming protocol specified in Section 6.1.1 is used in evaluating a contextualized semantic learning model where the information in an entire document is required. However, there are many different semantic learning models that do not consider the local context, e.g., all the models in learning global relatedness such as PCA and LSA as reviewed in Section 2.1. It seems unfair if we compare a contextualized semantic learning model to those that work only on a single term without access to the document-level information. To allow us to compare ours to more state-of-the-art semantic learning models, we extend the priming protocol defined in Section 6.1.1 by allowing all semantic learning model to use exactly the same information conveyed in an entire document in semantic priming. Hence, any model is provided with an entire document and the collective priming results of all the terms in this query document will be used for performance evaluation. In other words, the extended priming amounts to merging all the ranked lists achieved by different terms in the query document into a single document-level global ranked list with the same distance metric in the semantic representation space. For a term in the query document, however, priming itself actually results in a zero distance situation. If this result is allowed, almost all models can yield an error-free priming result. Therefore, we have to exclude such priming results of zero distance in our extended priming protocol. Given a query document  $\delta$  and a vocabulary  $\Gamma$ , the extended priming is defined by

$$\begin{aligned} E\_Prime(\delta) &= (\tau_i | \forall \tau_i, \tau_j \in \Gamma : \forall \hat{\tau} \in \delta \\ &\quad \wedge \hat{\tau} \neq \tau_i : \min \{e(\hat{\tau}, \tau_i|c)\} \\ &\quad \leq \min \{e(\hat{\tau}, \tau_j|c)\} \text{ if } i \leq j)_{i=1}^{|\Gamma|}. \end{aligned} \quad (10)$$

Eq. (10) results in a ranking list of  $|\Gamma|$  terms by using the minimum distance between any term  $\tau_i$  in  $\Gamma$  and all the  $|\delta|$  terms in a document  $\delta$ . As this protocol is designed for any semantic learning models no matter whether it uses the context or not, the distance measure  $e(\hat{\tau}, \tau_i|c)$  is decided by the nature of a semantic learning model; i.e.,  $c = \delta$  for a contextualized model and  $c = \text{null}$  otherwise. In Eq. (10), the condition “ $\forall \hat{\tau} \in \delta \wedge \hat{\tau} \neq \tau_i$ ” ensures that the zero distance information is never counted in finding out the minimum distance. Thus, this protocol guarantees all semantic learning models are fairly compared by performing document-level semantic priming with the same input and formulation in expressing their priming results.

### 6.1.3. Priming accuracy

In general, the priming performance is measured by the precision at  $K$  denoted by  $P@K$ ; i.e., the precision when only the top  $K$  entries in a ranked list are considered on a reasonable condition that  $K$  is less than the number of in-vocabulary terms  $|\Gamma|$ . Here, we denote the top  $K$  ( $K \leq |\Gamma|$ ) entries in a primed list by  $\text{Prime}_K(\tau, \delta) = (\tau_i)_{i=1}^K$  in the priming protocol (cf. Eq. (9)) or  $E\_Prime_K(\delta) = (\tau_i)_{i=1}^K$  in the extended priming protocol (cf. Eq. (10)). For a document  $\delta$ , the priming list achieved based on a priming protocol is thus defined by

$$\text{Prime}_K(x) = \begin{cases} \text{Prime}_K(\tau, \delta) \\ \quad x = (\tau, \delta) \text{ for the priming protocol} \\ E\_Prime_K(\delta) \\ \quad x = \delta \text{ for the extended priming protocol.} \end{cases}$$

Then  $P@K$  precision is defined as the ratio of primed terms in the ground-truth (i.e., all the terms in the query document  $\delta$ ) out of all  $K$  primed terms:

$$P@K(x) = \frac{|Prime_K(x) \cap \delta|}{K}.$$

Note that this measure is applicable to a query term in the priming or a query document in the extended priming protocols. For an evaluation dataset of multiple examples,  $X = \{x_i\}_{i=1}^{|X|}$ , the overall  $P@K$  precision is defined by

$$P@K(X) = \frac{\sum_{i=1}^{|X|} P@K(x_i)}{|X|}. \quad (11)$$

Intuitively, up to a prime level  $K$ ,  $P@K$  measures the precision of primed terms against the ground-truth to find out how many related terms appear in the top  $K$  primed terms. Due to the limitation of the ground-truth, the  $P@K$  measure may be affected by the cardinality of a query document, i.e.,  $|\delta|$ . In other words, only up to  $|\delta|$  primed terms can be confirmed definitely with the ground-truth. As  $K$  exceeds  $|\delta|$ ,  $P@K$  values might decrease rapidly for documents of few terms. Although  $P@K(X)$  may be a reasonable measure of comparison when  $K \leq |\delta|$ , it does not faithfully reflect the performance of any models when  $K > |\delta|$ . In contrast, the averaging precision on all the  $P@K(K = 1, \dots, |\delta|)$ , i.e.,  $AP(x) = \frac{\sum_{K=1}^{|\delta|} P@K(x)}{|\delta|}$ , automatically adapts for the various lengths of different documents used as ground-truth by only concerning the top  $|\delta|$  entries of the primed list resulting from a query instance, which provides a reliable performance measure. The overall average precision on a test dataset of  $|X|$  examples is

$$MAP(X) = \frac{\sum_{i=1}^{|X|} AP(x_i)}{|X|}. \quad (12)$$

In essence, semantic priming in response to a query concept is an information retrieval task and hence the evaluation measures commonly used in information retrieval are applicable. The Area Under Curve (AUC) is a commonly used measure by calculating the area formed under the curve of precision as a function of recall at the standard 11 recall levels:  $\mathbb{I} = \{0.0, 0.1, \dots, 1.0\}$  (Manning, Raghavan, & Schütze, 2008, pp. 158–163). Precision and recall at a specific recall level  $l \in \mathbb{I}$  are:

$$Precision(\ell|x) = \frac{|Prime_K(x) \cap \delta|}{K},$$

where  $\ell = Recall(K|x) = \frac{|Prime_K(x) \cap \delta|}{|\delta|}$ .

$Recall(K|x)$  specifies a certain recall level  $l = k/|\delta|$  implying that at least  $k$  out of all the  $|\delta|$  related terms in the ranked list have been retrieved and is used to form the measure  $Precision(\ell|x)$ , i.e., precision  $P@K$  at level  $l$ . Accumulating the precision values at all the 11 recall levels across an evaluation dataset leads to an overall AUC measure:

$$Precision(\ell|X) = \frac{\sum_{i=1}^{|X|} Precision(\ell|x_i)}{|X|},$$

$\ell = 0.0, 0.1, \dots, 1.0$ . (13)

Intuitively, a larger AUC region formed by  $Precision(\ell|X)$  suggests that more of the related terms have been retrieved at the standard recall levels and the precision–recall curve clearly shows the performance of a tested model at different recall levels.

In summary, four criteria,  $P@K$ ,  $MAP$ ,  $Precision/Recall$  and  $AUC$ , are used in our experiments to evaluate the priming/extended priming performance of a semantic learning model.

## 6.2. Experimental protocols

For a thorough performance evaluation in semantic priming, we have designed a number of experiments in different settings corresponding to several real scenarios to test the learnt semantic representations, including: (a) domain-specific semantics: test on all the documents used in training a model and those unseen documents in the same corpus; i.e., a subset of documents were not used in training; (b) transferability: test on the different corpora where none of documents in those corpora were used in training; (c) noisy data: test on incomplete local context; (d) OOV data: testing on synthesized and real documents of OOV terms; and (e) Comparison: comparing ours to those semantic learning models reviewed in Section 2 with exactly the same settings. As described in Section 5.2, we conducted the cross-validation in training a semantic model for three trials. Hence, the averaging accuracy along with standard error arising from three trials is reported in terms of two priming protocols described above.

### 6.2.1. Within-corpus setting

The within-corpus setting refers to the evaluation that uses the training or the test documents subsets from the training corpora (cf. Section 5.1); i.e., CAL500, MagTag5K and Corel5K. The use of training documents in this setting is expected to test the quality of semantics learnt by a model in terms of this application. Moreover, measuring the priming accuracy on the train documents mimics a real scenario where all available information is used in building up a semantic space to be used in a variety of applications later on. On the other hand, by using the test document subset in this setting, we would evaluate the generalization of the learnt semantics into unseen documents that were probably annotated by the same cohort of users. We refer to such evaluation as within-corpus test (WCT) and expect that this setting would examine the quality and the generalization of learnt semantics in a domain-specific sense.

### 6.2.2. Cross-corpora setting

Unlike the WCT, we design experiments to test unseen documents in corpora that were never used in semantic learning. We refer to this type of evaluation as cross-corpora test (CCT). As a result, the CCT would investigate the transferability of learnt semantics in terms of this application. Table 2 summarizes our cross-corpora settings including training and test datasets, the number of common terms shared by training and test datasets, the number of documents of shared terms and the number of documents containing OOV terms in a test dataset.

### 6.2.3. Incomplete local context setting

To achieve a contextualized semantic representation of a term, both the term and its local context, i.e., all accompany terms in the document containing it, must be required as described in our problem formulation in Section 1. In real applications, a test document could mismatch training data. For instance, it could be a subset of a training document using fewer yet more informative terms or an enhanced version of a training document by adding more terms. In this setting, we would design experiments to test mismatched documents. As argued in Section 6.1.1, it does not seem possible to attain all the concepts and their similarities in terms of all possible contexts. Thus, it is impossible for us to simulate on one mismatch situation that more terms are added to existing documents. Fortunately, we can simulate the other mismatch situation, incomplete local context, by removing a few terms from existing documents.

To simulate the incomplete local context situation, a training example of complete local context,  $x = (\mathbf{t}(\tau), \mathbf{I}(\tau|\delta))$ , is altered

**Table 2**  
Summary of cross corpora (CC) experimental setting.

Training dataset	Test dataset	#Common term	#CC document	#CC-OVV document
MagTag5K	MSD	75	817	39,507
Corel5K	LabelMe	105	520	8,703
Corel5K	SUNDatabase	90	266	11,935

into a corrupted version,  $\tilde{x} = (\mathbf{t}(\tau), \mathbf{I}(\tau|\tilde{\delta}))$  where  $\tilde{\delta}$  is a subset of the original document  $\delta$  achieved by removing a number of terms randomly from  $\delta$ . The incomplete context,  $\mathbf{I}(\tau|\tilde{\delta})$ , corresponds to the topics distribution obtained from the incomplete document  $\tilde{\delta}$ . As a result, the use of fewer accompany terms in  $\tilde{\delta}$  results in larger uncertainty in semantic priming and hence causes a bigger difficulty in priming all the accompany terms in the original document  $\delta$ . Here, we emphasize that the ground-truth is the original document but the local context is derived from a subset of this document in semantic priming under this setting. In our incomplete local context experiments, we used the missing rate defined by  $(1 - \frac{|\tilde{\delta}|}{|\delta|}) * 100\%$  to control the number of terms removed randomly from a complete document. In this paper, we report results based on the missing rate in different ranges: up to 10% and between 10% and 30% due to the variable length of different documents.

#### 6.2.4. Out of vocabulary (OOV) setting

The OOV problem appears challenging in semantic learning from descriptive terms. Based on our proposed approach, we have proposed two methods to deal with OOV terms as described in Section 4.4. Here, we would use semantic priming to evaluate our proposed methods.

In our OOV experiments, we used the reserved subset of MagTag5K as described in Section 5.1. In this reserved subset, there are 1160 documents where each of them contains at least one out of the 22 reserved terms used as simulated OOV terms. Their concept **CE** representations achieved from the semantic model trained on MagTag5K were used in semantic priming. Moreover, we also used the real documents containing OOV terms in the test corpora (cf. our CCT setting in Section 6.2.2). As a result, there are 39,507 documents involving 23,619 OOV terms in the MSD, 8703 documents of OOV 2110 terms in the LabelMe and 11,935 documents containing 2068 OOV terms in the SUNDatabase used in our OOV experiments. For those OOV documents in the MSD, the semantic model trained on MagTag5K was used to generate their **CE** representations. For those OOV documents in the SUNDatabase and the LabelMe, the semantic model trained on Corel5K was employed to yield their **CE** representations. The information on the cross-corpora OOV setting is also listed in Table 2.

To the best of our knowledge, those approaches used in our comparative studies do not address the OOV issue. Hence, the OOV experiments only involve our proposed approach described in Section 4.4 and the priming protocol is only employed for performance evaluation.

#### 6.2.5. Comparison settings

We use the learning models reviewed in Section 2 as baselines for comparative studies. For training and test, we apply the exact same cross-validation protocol described in Section 5.2 to each semantic learning model. As a result, the information on training those models is summarized as follows:

**Latent Semantic Analysis (LSA):** As described in Section 2, the unsupervised dimensionality reduction technique is performed using the training documents and model selection was done by using the percentage of variance (POV) measure by monitoring eigenvalues  $\lambda_i$  resulting from the matrix decomposition. As a

result,  $n$  features are employed when the top  $n$  eigenvalues cover at least 90% of the variance of training data; i.e.,  $POV = (\sum_{i=1}^n \lambda_i) / (\sum_{i=1}^{|V|} \lambda_i) \geq 90\%$ . As a result, we retained 35, 25 and 80 features for MagTag5K, CAL500 and Corel5K, respectively. The same numbers of features were extracted for test data.

**Principle Component Analysis (PCA):** PCA relies on preprocessing and aggregation of the document-term binary matrix followed by dimensionality reduction of the aggregated matrix in order to obtain per term feature vectors. In our experiments, we applied preprocessing techniques including: using the binary term frequency, the *tfidf* re-weighting and Positive Point-wise Mutual Information (PPMI) re-weighting as preprocessing. Also we considered different distance metrics in measuring the term-to-term relatedness such as the cosine, the co-occurrence (non-normalized cosine), Kullback–Leibler divergence and Hellinger divergence as aggregation measures. For each of those combinations of preprocessing and aggregation, we performed unsupervised dimensionality reduction and evaluated the resultant semantic space by using documents from MagTag5K and CAL500 based on their  $P@2$  priming performance. The combination that produced the best results is the *tfidf* reweighed matrix followed by the co-occurrence aggregation measure. As a result, we shall report results based on this combination.

**Information Theoretic Smoothing (InfoTheo):** This model started with smoothing the binary BoW representation of each document by using information regarding the pairwise use of terms over an entire training dataset. Following the suggestions in Mandel et al. (2011), we obtained this pairwise information by using all the 12 aggregation methods listed in the PCA setting and applied such information to the smooth document-term matrix generation. This matrix requires a further tuning of the two parameters, the number of associated terms  $k$  and the reweight factor  $\alpha$ . With the suggestions in Mandel et al. (2011), we tuned those parameters by a grid search on a reasonable range for each of training datasets with different aggregation methods, respectively. We looked into the situations as  $k = 1, 3$  and 5 terms while reweighting the matrix using different factors for  $\alpha = 0.1, \dots, 0.5$ . A total of 180 experiments were carried out in each of MagTag5K and CAL500. We observed that the *tfidf* reweighed matrix followed by the co-occurrence aggregation measure performed significantly better than other 11 aggregation methods. As a result, we applied the best aggregation method to MagTag5K, CAL500 and Corel5K. We report results based on the setting corresponding to the best  $P@2$  training performance in this paper. In detail, the optimal parameters are  $k = 1, \alpha = 0.2$  for MagTag5K,  $k = 1, \alpha = 0.1$  for CAL500 and  $k = 3, \alpha = 0.3$  for Corel5K. Furthermore, we conducted experiments by using a full term-to-term matrix and PCA dimensionality reduced version of this matrix. We observed that the dimensionality reduced version generally outperforms the full matrix on different datasets. Also this processing allows both InfoTheo and other global relatedness learning model to have the same dimension in their representation spaces. In this paper, we report only the results generated by the dimensionality reduced version.

**Skip Gram:** In order to avoid capturing any unreal syntactic structure, we randomize the order of terms in each document before the Skip Gram learning. Training a Skip Gram model requires tuning two hyper-parameters: dimension of the embedding space



and size of the neighborhood window used to specify the context. Using a grid search, we trained a number of Skip Gram models for a training dataset and selected the one with the best  $P@2$  training performance to report their results in this paper. We observed that the performance on different datasets was not sensitive to the dimensionality of the embedding space but affected by the window size. In general, the smaller the window size, the better the model performed. As a result, we selected the models that had the window sizes of one, three and one for MagTag5K, CAL500 and Corel5K, respectively, and the dimension of the embedding space on the three datasets is the same used for PCA, i.e., 35, 25 and 80 for MagTag5K, CAL500 and Corel5K, respectively. In our experiments, we use the word2vec code (Mikolov et al., 2013) to train the Skip Gram models.

**Latent Dirichlet Allocation (LDA) and Probabilistic LSA (PLSA):** The same number of features used in representing our local context was employed for the LDA evaluation; i.e., 19, 25 and 20 features for MagTag5K, CAL500 and Corel5K, respectively. The hyper-parameter tuning was described in Section 5.2. For LDA, we used the standard C implementation of LDA (Blei et al., 2003) in our experiments. The same number of topics used in LDA was adopted for the PLSA as the unique difference in the two methods is the used distributions in capturing document-level semantics. A PLSA model was trained using the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977) with convergence of the likelihood as a stopping criterion.

**CRBM:** The use of binomial units in CRBM requires tuning the number of units in the hidden layer. We conducted a number of experiments with different latent space dimensions and observed that the performance was insensitive to the dimensionality of latent space. This can be explained by the fact that the CRBM is designed to smooth the term-to-document relatedness rather than term-to-term relatedness. In our experiments, we used the same number of hidden units in CRBM as that used in our CE space. CRBM models were trained with the contrastive divergence algorithm (Hinton, 2002) where we used the recommended learning rate of 0.1 with moment 0.5. Our implementation is based on the MatRBM<sup>1</sup> package.

**Random:** This is a model used to form a baseline without learning. Depending on an evaluation criterion, the model worked by returning a proper number of terms uniformly sampled from a test dataset to form the primed list for a given query term.

Once those models were trained, the following methods were used in semantic priming as well as ranking the different terms for a query instance or a query document:

**PCA, LSA, InfoTheo and Skip Gram:** We have investigated two distance metrics in our priming experiments, i.e., the cosine and the Euclidean distances. As the cosine metric outperformed Euclidean for all models, we used the cosine metric to measure the distance between different terms in the semantic representation space.

**LDA and PLSA:** The information theoretic distance between a pair of terms given a topic distribution was used (cf. Section 2.2).

**CRBM:** The model was tested with 100 trials by using one-hot representation of the query term, i.e., a vector with all zeros except one unit corresponding to the query term set to one, and one-hot representation of the query document as context. In each trial, the model acted 100 forward and backward steps. The resultant 100 output vectors on the visible units achieved are averaged and the averaged output of visible units was used to measure relatedness; for the visible units of the higher activation values in output, their corresponding terms are treated as having higher relatedness. Note that due to the technical limitation of this model (cf. Sections 2.3

and 7), it could be evaluated only on training sets in the WCT experiments.

**Random:** The model ranks the terms randomly with a uniform distribution.

**Our Model:** The terms are ranked based on their Euclidean distances in the concept embedding (CE) space to a query concept. It should be clarified that the prediction learning in our model may lead to a CE space that facilitates the final CE space formation via distance learning (cf. Section 4). To evaluate the gain of distance learning with our proposed Siamese architecture, we apply the **CE** representations achieved via the initial prediction learning, named **CE**, and the final distance learning, dubbed **Siamese-CE**, to semantic priming.

In summary, our comparative studies in applying different semantic learning models in semantic priming are based on exactly the same experimental settings. While all of the aforementioned semantic learning models were evaluated with the extended priming protocol, only LDA, PLSA, CRBM and Random models along with ours were evaluated in the priming protocol described in Section 6.1 since only these models can generate their semantic representations with a term and its local context simultaneously.

### 6.3. Within-corpus results

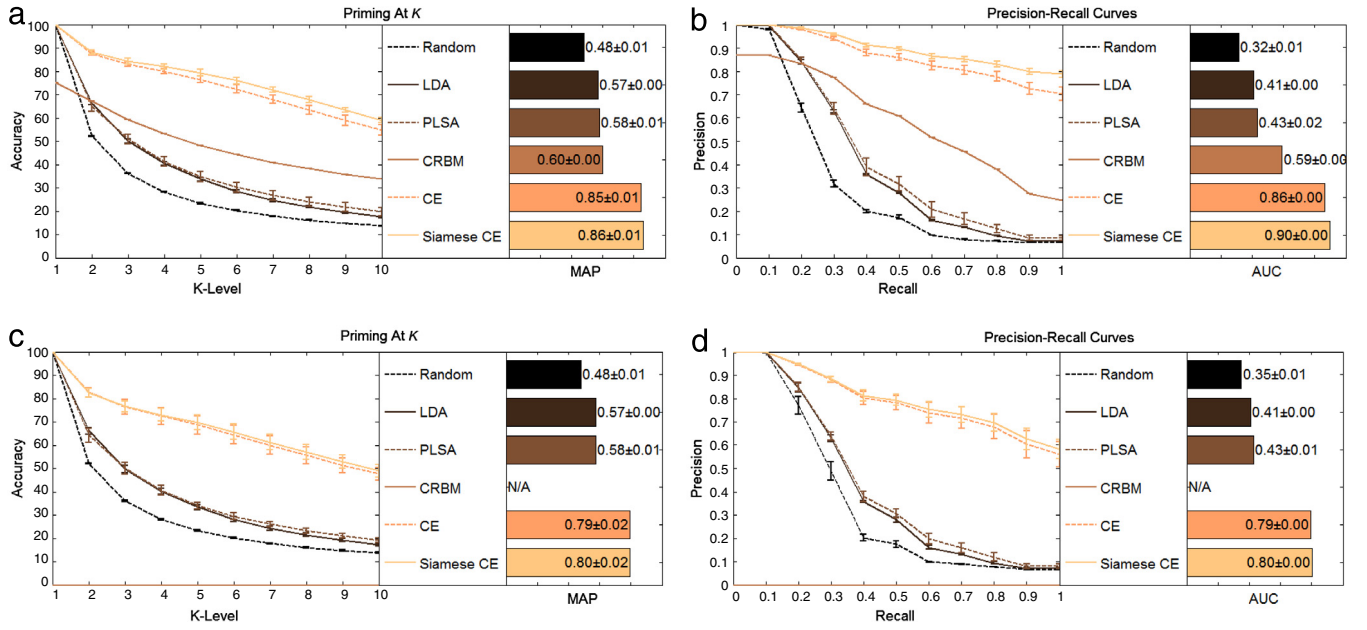
With the experimental setting described in Section 6.2.1, we report the WCT experimental results on three training corpora: MagTag5K, CAL500 and Corel5K in terms of two priming protocols.

#### 6.3.1. Priming results

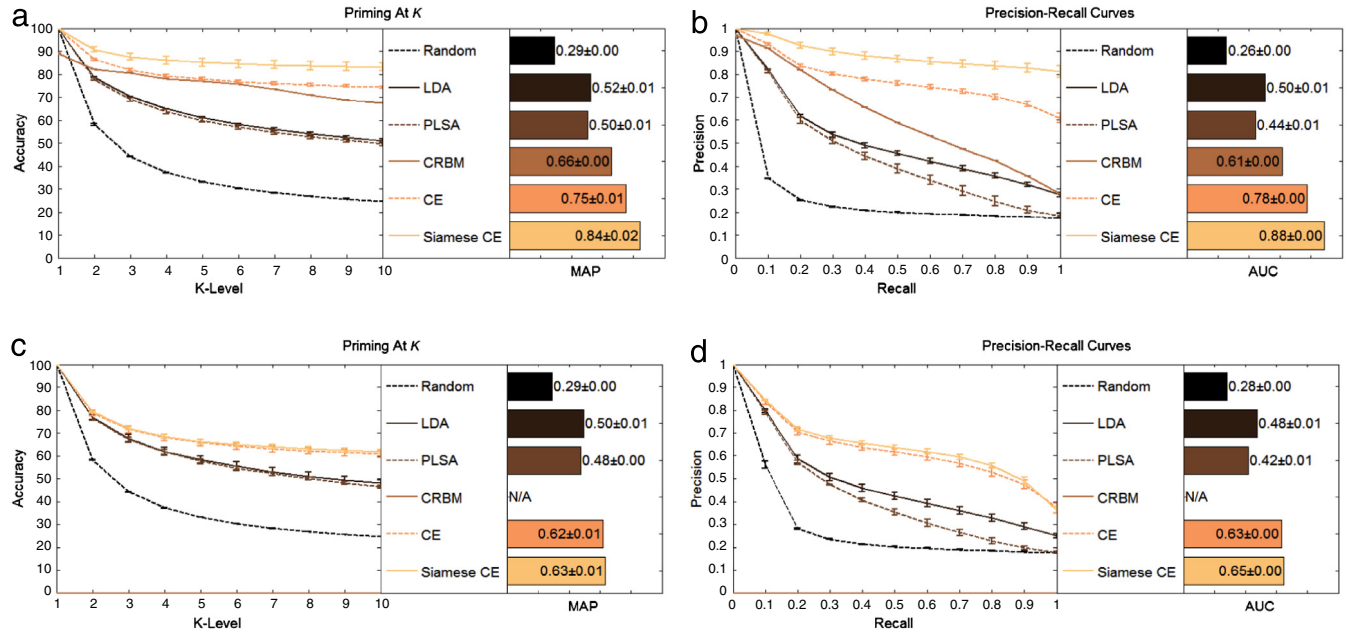
Fig. 5 illustrates the priming results of five different models on MagTag5K in terms of four evaluation criteria defined in Section 6.1.3. Fig. 5(a) shows the priming results on the training subset in terms of  $P@K$  as  $K$  varies from one to 10 and the MAP results, both indicated by the mean and standard error on three trials. It is observed from Fig. 5(a) that all the models apart from CRBM outperform the Random model regardless of  $K$  and the length of evaluated documents. The CRBM performs worse at  $K = 1$  but much better than the Random for different  $K$  values up to 10 in terms of  $P@K$  and the MAP. Normally, the ground-truth for  $P@1$  corresponds to the query term itself and the stochastic nature of CRBM might be responsible for its failure at  $K = 1$ . Fig. 5(b) shows the priming results on the training subset in terms of the precision–recall performance at 11 standard recall levels and the aggregated AUC. The same as seen in Fig. 5(a) is observed. It is evident from Fig. 5(a) and (b) that our model performs the best among all five models regardless of evaluation criteria. In particular, Siamese-CE leads to the significantly better performance by beating the runner-up, the CRBM, with a big margin, e.g., 26% in MAP and 31% in AUC. Also we observe that Siamese-CE performs slightly better than CE on the training subset. Fig. 5(c) and (d) show the priming results on the test subset in terms of four performance indexes, respectively. The exactly same as seen on the training subset is observed on the test subset although the performance of all the models on the test subset is degraded in comparison to that on the training subset. While the CRBM is no longer applicable to the test subset, Siamese-CE still wins with a big margin of at least 22% in MAP and at least 37% in AUC in comparison to other three models. It is also observed that **CE** representation seems to have a better generalization capability than Siamese-CE although Siamese-CE still performs better than CE on the test subset. Overall, our model outperforms others with the statistical significance ( $p$ -value  $< 0.01$ , Student's  $t$ -test) apart from  $K = 1$ . The experimental results on this dataset demonstrate that the accurate concepts and their relatedness have been captured by using both terms and their local context and such

<sup>1</sup> <https://code.google.com/p/matrbm/>.





**Fig. 5.** Priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC on MagTag5K. (a–b) Training subset. (c–d) Test subset. Error bars regarding the  $P@K$  and the Precision vs. Recall curves indicate standard error and the numbers regarding the MAP and the AUC are mean and standard error of the MAP and AUC priming accuracy. This notation is applied to all the figures hereinafter.



**Fig. 6.** Priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC on CAL500. (a–b) Training subset. (c–d) Test subset.

learnt semantics can be well generalized to those documents that were never seen in training.

Fig. 6 shows the priming results of five different models on CAL500 in terms of four evaluation criteria. It is observed from Fig. 6(a) and (b) that our model performs significantly better than other models on the training subset given the fact that Siamese-CE yields at least 18% in MAP and at least 27% in AUC higher accuracy than other models. It is also observed that the high document cardinality of this dataset makes the Random model relatively easy to guess a few related terms, i.e., results in relatively high  $P@K$  for small  $K$  values, as evident in Fig. 6(a). In Fig. 6(b), it is seen that higher precision at high recall levels is achieved than that achieved on the training subset in MagTag5K. As a runner-up, however, the performance of CRBM decreases rapidly as the

recall level increases. This suggests that the CRBM had encountered a difficulty in identifying all the terms related to a query concept. The same problem can be found in other models except ours. Fig. 6(c) and (d) illustrate the performance of different models on the test subset. Overall, the same conclusions drawn on the training subset are reached on the test subset; Siamese-CE yields the statistically significant better performance ( $p$ -value  $< 0.01$ , Student's  $t$ -test) than other models by winning at least 13% in MAP and at least 17% in AUC on the test subset. In comparison to the results on MagTag5K shown in Fig. 5, our model generally behaves consistently though the generalization performance on CAL500 is worse than that on MagTag5K. As described in Section 5.1, CAL500 is a music tag collection quite different from MagTag5K in terms of length of documents or document cardinality and the tag usage

distribution (cf. Table 1). In light of capturing the accurate concepts and their relatedness, the experimental results on two distinct music datasets suggest that our model is not sensitive to document cardinality and statistics underlying different collections in the same domain as is evident in Figs. 5 and 6.

Fig. 7 illustrates the priming results of five different models on Corel5K in the image domain in terms of four evaluation criteria. Overall, our model yields the statistically significant better results ( $p$ -value  $< 0.01$ , Student's  $t$ -test) than other models. On the training subset, it is evident from Figs. 7(a) and 8(b) that Siamese-CE leads to at least 12% in MAP and at least 24% in AUC higher than others. From Fig. 7(c) and (d), the favorable generalization capability of our model is seen clearly; on the test subset, Siamese-CE considerably outperforms other models by winning at least 12% in MAP and 33% in AUC. In addition, the gain of Siamese-CE over CE is more visible on this dataset. From Fig. 7, however, it is also observed that the performance is degrading rapidly across the ranked list due to the nature of this dataset. As described in Section 5.1, a document in this dataset contains only five labels at maximum and 3.5 labels on average, but there is a vocabulary of 292 different labels in this dataset. Once the value of  $K$  in  $P@K$  and the recall level reach a certain degree beyond the length of a query document, the performance is inevitably degraded regardless of which model is used. Even in this situation, the experimental results shown in Fig. 7 suggest that our model still yields the significantly better performance, in particular, at high recall levels, as is evident in Fig. 7(b) and (d). In general, the results on this dataset demonstrate the capability of our model in capturing the accurate concepts and their relatedness from documents containing a small number of terms.

### 6.3.2. Extended priming results

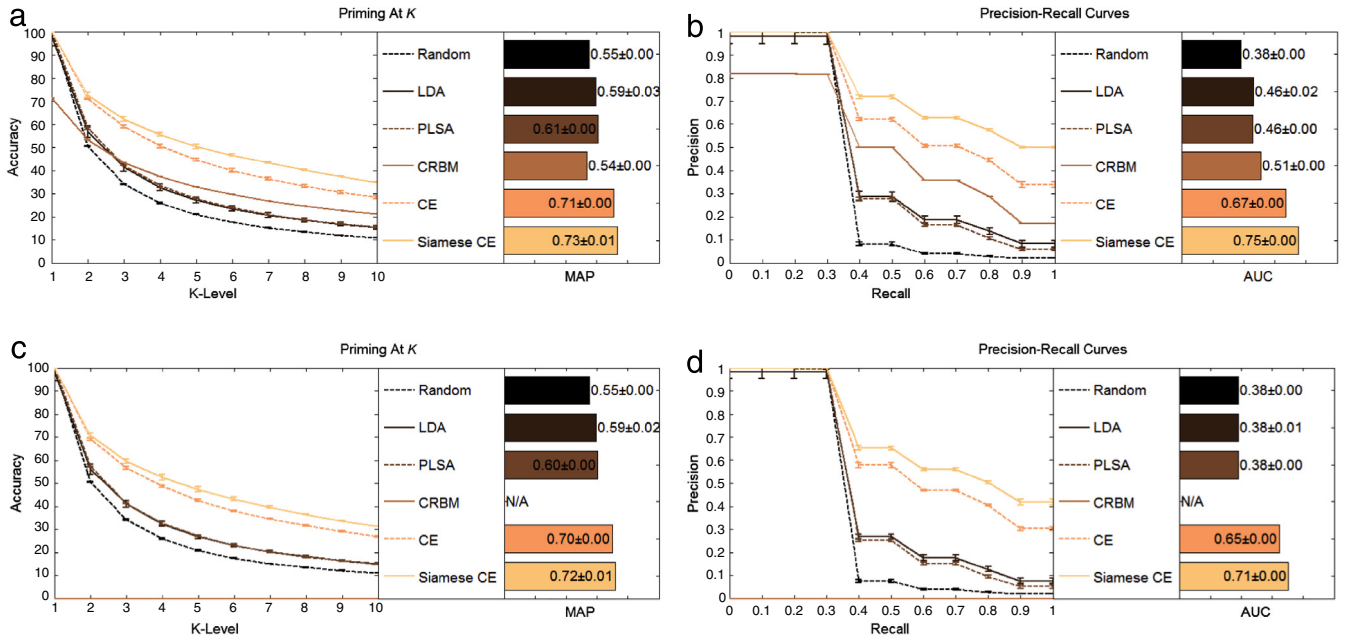
Fig. 8 illustrates the extended priming results of nine different models on MagTag5K in terms of four evaluation criteria defined in Section 6.1.3. Regarding the results on the training subset shown in Fig. 8(a) and (b), our model always outperforms all other models with the statistical significance ( $p$ -value  $< 0.01$ , Student's  $t$ -test) in all four evaluation criteria. In particular, Siamese-CE wins at least 26% in MAP and 34% in AUC over other models. As shown in Fig. 8(c) and (d), our model also performs the best on the test subset, and moreover, Siamese-CE beats the runner-up with a big margin of 23% in MAP and 22% in AUC. On both training and test subsets, our model performs particularly well at high recall intervals as shown in Fig. 8(b) and (d). Overall, CE performs equally well on both training and test subsets in MagTag5K. A closer look suggests that Siamese-CE outperforms CE at high recall levels on the training set but this advantage disappears on the test subset. The better performance achieved by Siamese-CE on the training subset is thanks to the distance learning that refines the CE representation. On both training and test subsets, LSA, Skip Gram, LDA and PLSA all perform poorly although they win over the Random model. Interestingly, LDA and PLSA are two probabilistic topic models (PTMs) that yield document-level representations. In this document-level priming evaluation, however, the PTMs do not seem to be able to capture the subtle difference in the concepts conveyed in a query document, which provides evidence to support our contextualized semantic learning problem formulation. From Fig. 8, it is also evident that Skip Gram cannot capture the semantics from tags well due to a lack of syntactic context in documents of descriptive terms. In contrast, the non-contextualized models, PCA and InfoTheo, perform well given the fact they win over almost all other models apart from ours as illustrated in Fig. 8. On the training subset, however, CRBM performs better than PCA and InfoTheo at both small  $K$  in  $P@K$  and low recall intervals, as shown in Fig. 8(a) and (b), due to its capability in capturing document-term relatedness. It is worth

stating that the success of PCA and InfoTheo relies on the careful weighting of the document-term matrix and proper aggregation and those results reported here are those corresponding to the optimal parameters. Finally, the results on MagTag5K in both priming and extended priming shown in Figs. 5 and 8 also raise an issue on why the distance learning by our Siamese architecture does not lead to a substantial gain on this dataset, in particular, regarding generalization, which will be discussed later on.

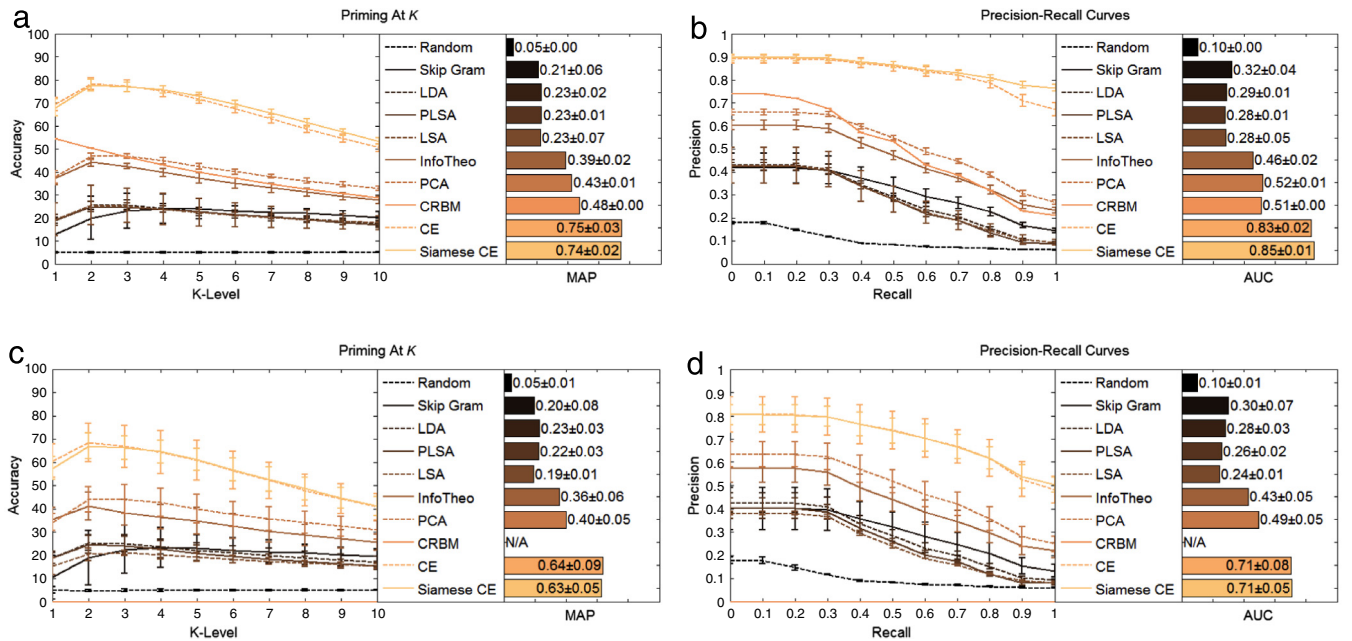
Fig. 9 shows the extended priming results of nine different models on CAL500 in terms of four evaluation criteria. Once again, our model outperforms other models regardless of evaluation criteria. As shown in Fig. 9(a) and (b), the results on the training subset indicate that Siamese-CE wins over other models at least 18% in MAP and at least 25% in AUC and, in particular, our model performs much better at high recall levels. In comparison to results on MagTag5K, there are two non-trivial changes: CRBM outperforms PCA and InfoTheo considerably and Siamese-CE performs significantly better than CE on the training subset of this dataset. Nevertheless, the results on the test subset shown in Fig. 9(c) and (d) reveal that all the models including ours seem to face difficulty in extended priming especially at high recall levels. The difficulty causes the performance of some models to be close to that of the Random model. An analysis on the training subset reveals that it may be caused by a lack of sufficient informative training examples given the fact that 335 training documents actually consist of 158 different tags. Due to insufficient training data reflecting various concepts and intended terms' use patterns, it is likely that the learning may overfit the training data and hence some unseen positive instances may be grouped incorrectly with negative instances in our distance learning. Consequently, our winning margin over other models becomes smaller in comparison to results on MagTag5K (cf. Fig. 8(a) and (b)), e.g., Siamese-CE gains only 5% in MAP and nothing in AUC in comparison to the runner-up, PCA. As a non-contextualized model, PCA learns the global relatedness of tags. In the presence of insufficient training documents for capturing the accurate concepts, the PCA may be a choice after trade-off between performance gain and computational efficiency in this document-level retrieval task.

Fig. 10 illustrates the extended priming results of nine different models on Corel5K in the image domain in terms of four evaluation criteria. Overall, our model yields the statistically significant better performance ( $p$ -value  $< 0.01$ , Student's  $t$ -test) than other models on both training and test subsets. As shown in Fig. 10(a) and (b), the results suggest that Siamese-CE wins over other models at least 10% in MAP and at least 20% in AUC on the training subset. Once again, non-contextualized models, PCA and InfoTheo, outperform other models apart from ours. Fig. 10(c) and (d) illustrate the results on the test subset where all the models are ranked as same as done on the training subset in terms of their performance. Siamese-CE wins over the runner-up, InfoTheo, 9% in MAP and 15% in AUC and also leads to better generalization than CE with the gain of 5% in MAP and 9% in AUC. In particular, Siamese-CE outperforms CE at high recall levels in both training and test subsets as shown in Fig. 10(a) and (b). It indicates that the distance learning would be paid off should there be sufficient informative training examples regarding various concepts and intended term-use patterns. Once again, LDA and PLSA perform poorly on this dataset as is evident in Fig. 10, which lends further evidence to support our contextualized semantic learning given the fact that a huge gain is brought by our model based on LDA.

In summary, the WCT experimental results on different data sets in two different priming protocols demonstrate that our approach generally outperforms other state-of-the-art methods in semantic priming and has the proven generalization capability that the learnt semantics can be applied to unseen documents in training for this retrieval task.



**Fig. 7.** Priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC on Corel5K. (a–b) Training subset. (c–d) Test subset.



**Fig. 8.** Extended priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC on MagTag5K. (a–b) Training subset. (c–d) Test subset.

#### 6.4. Cross-corpora results

In the CCT experiments, we apply the semantic representation achieved by a model trained on a corpus to another test collection for semantic priming. Here, we report results for those semantics trained on MagTag5K and applied to MSD as well as those trained on Corel5K and applied to LabelMe and SUNDatabase in terms of two priming protocols.

##### 6.4.1. Priming results

Fig. 11 illustrates the priming results of four different models on three test collections in terms of four evaluation criteria defined in Section 6.1.3.

Fig. 11(a) and (b) show the priming results on MSD. Overall, our model outperforms other model with the statistical significance

( $p$ -value  $< 0.01$ , Student's  $t$ -test). It is observed from Fig. 11(a) and (b) that Siamese-CE gains at least 9% in MAP and at least 19% in AUC higher than other models and, in particular, yields the considerably better performance at high recall levels. Also CE leads to a considerably better performance than other models and its performance is slightly lower than that of Siamese-CE. In contrast, LDA and PLSA yield the results close to those generated by the Random model, which indicates the poor transferability of semantics learnt by two models. By comparison to the results on the test subset of MagTag5K shown in Fig. 5(c) and (d), we observe that the performance of Siamese-CE on MSD is worse than the WCT results, e.g., 12% in MAP and 14% in AUC lower. As MagTag5K is a small subset of MSD, it is likely that there are much more varied patterns and concepts associated with a tag in MSD and different annotators working on the large collection,



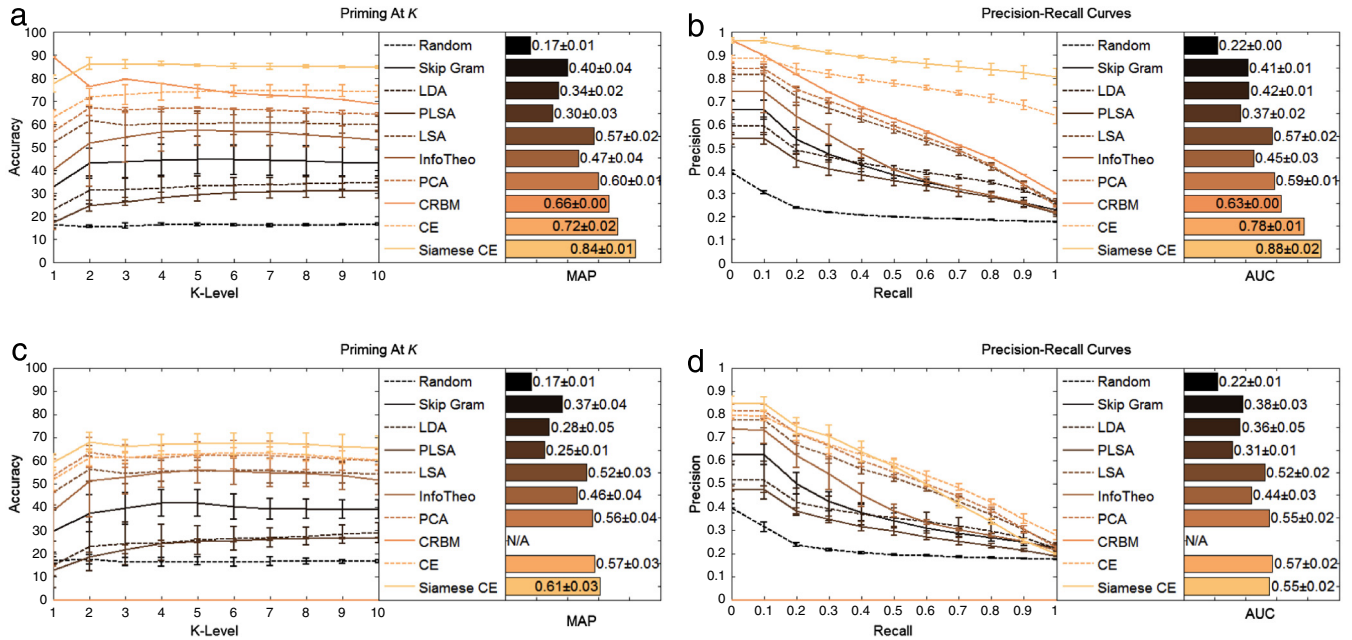


Fig. 9. Extended priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC on CAL500. (a–b) Training subset. (c–d) Test subset.

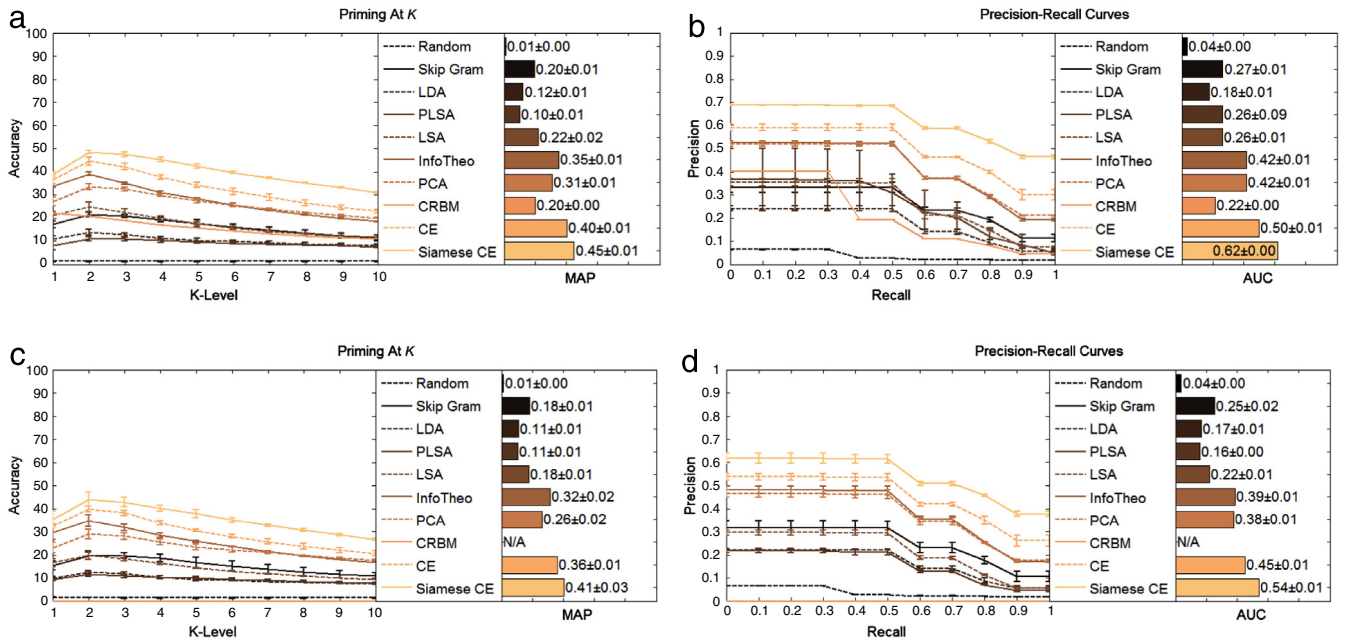


Fig. 10. Extended priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC on Corel5K. (a–b) Training subset. (c–d) Test subset.

which could have more and alternative interpretations for those tags in MagTag5K in a much larger tag vocabulary in MSD. Despite the degraded performance on MSD, we believe that the priming results generated by our model are quite promising in learning transferable semantics.

Fig. 11(c) and (d) show the priming results on LabelMe. It is observed that our model outperforms other model with the statistical significance ( $p$ -value  $< 0.01$ , Student's  $t$ -test); Siamese-CE wins over the runner-up 12% in MAP and 27% in AUC and displays significantly better performance at high recall levels. Also the performance of CE is superior to that of other models but lower than Siamese-CE. Unfortunately, LDA and PLSA yield poor performance, roughly identical to the Random model, as clearly seen in Fig. 11(c) and (d). In contrast to the results on the test subset of Corel5K shown in Fig. 7(c) and (d), it is observed that the

performance of Siamese-CE on LabelMe is close to the WCT results, e.g., only 4% in MAP and 9% in AUC lower. Also it still maintains the good performance at high recall levels. Those results suggest quite strongly that our model can capture the transferable semantics when training and test corpora have a high agreement in intended meanings of terms in annotation.

Fig. 11(e) and (f) show the priming results on SUNDatabase. Once again, our model outperforms other model with the statistical significance ( $p$ -value  $< 0.01$ , Student's  $t$ -test); Siamese-CE wins over the runner-up 13% in MAP and 25% in AUC and, in particular, the significantly better performance at high recall levels. It is also observed that all the models perform on this dataset very similarly to those on LabelMe although Siamese-CE yields a lower priming accuracy in comparison to that on LabelMe, e.g., 3% in both MAP and AUC.



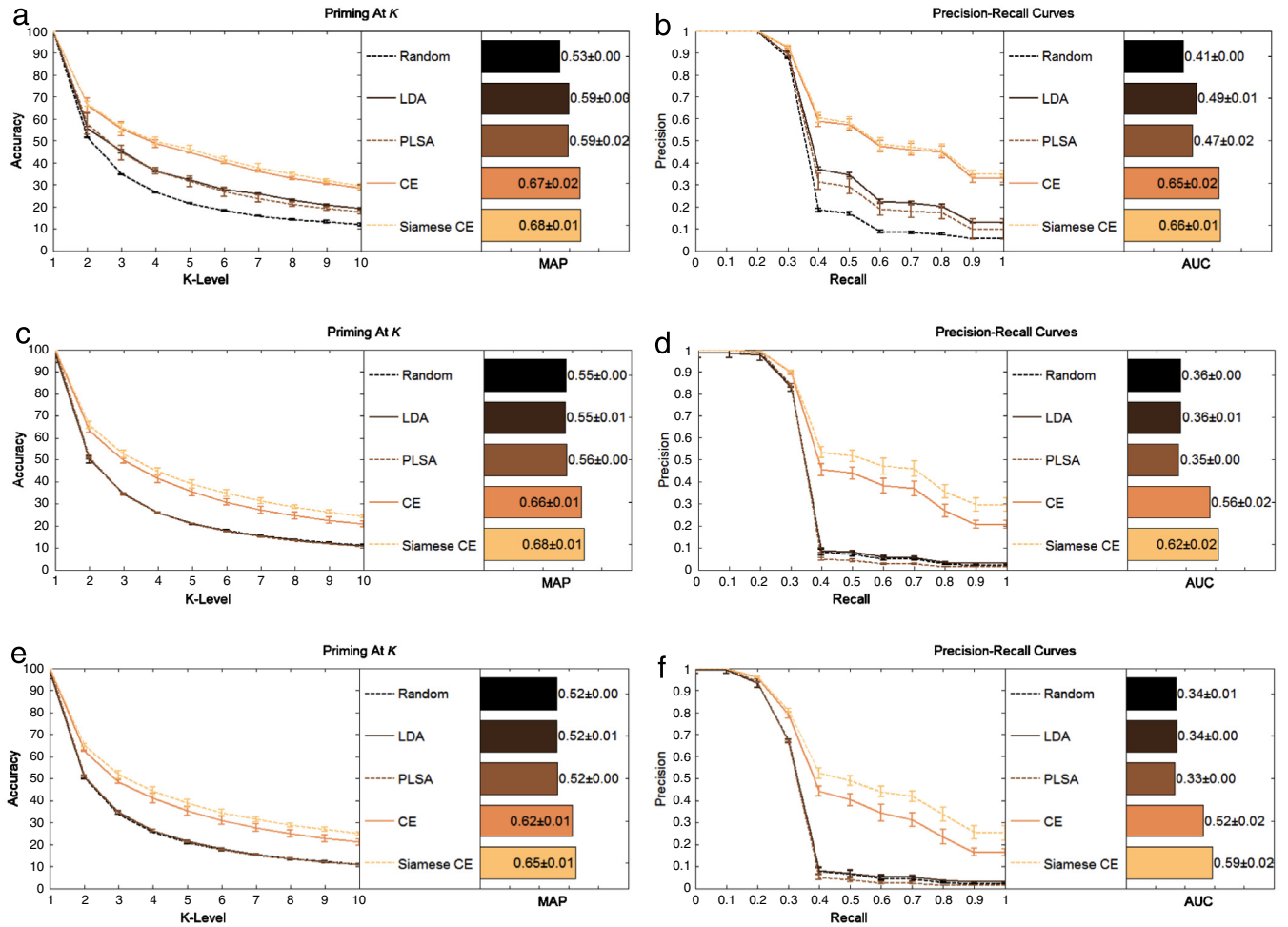


Fig. 11. Priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC in the CCT experiments. (a–b) MSD. (c–d) LabelMe. (e–f) SUNDatabase.

For three image datasets, we notice that the document cardinality is quite different given the fact that on average there are 3.5, 7.3 and 11 labels per document in Corel5K, LabelMe and SUNDatabase, respectively. This information implies that our model is less sensitive to some statistical variation but more sensitive to the semantics underlying co-occurring terms.

#### 6.4.2. Extended priming results

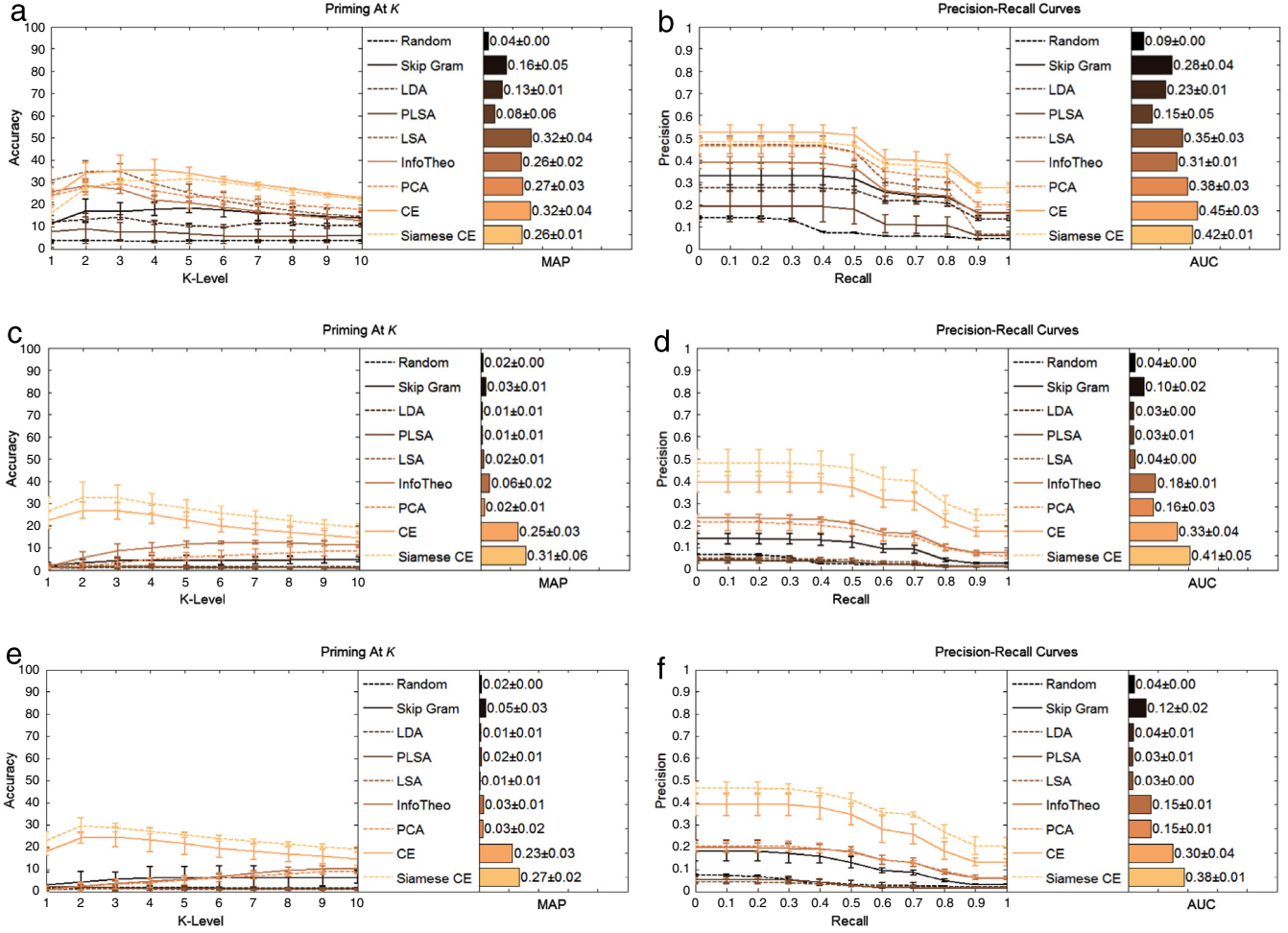
Fig. 12 shows the extended priming results of nine different models on three test collections in terms of four evaluation criteria defined in Section 6.1.3.

Fig. 12(a) and (b) illustrate the extended priming results on MSD. In comparison to the performance on the test subset on MagTag5K shown in Fig. 8(c) and (d), all the models including ours perform poorly, which demonstrates the challenge in learning transferable semantics with limited training data. It is observed that LSA performs the best in MAP while our model wins in AUC. In general, our model performs better at large  $K$  and high recall levels while LSA outperforms others at small  $K$  and low recall levels. In particular, CE always outperforms Siamese-CE. For the reason described in Section 6.4.1, a contextualized model is more sensitive to the usage patterns and intended meanings of terms in capturing concepts in context than a non-contextualized model that learns only global relatedness. In general, both the priming and the extended priming results on MSD suggest that a contextualized semantic model does not seem to transfer the semantics learnt from a less informative dataset to those of richer information, intricate concepts and alternative intended term-use patterns.

Fig. 12(c) and (d) show the extended priming results on LabelMe. It is observed that our model outperforms other models with the statistical significance ( $p$ -value  $< 0.01$ , Student's  $t$ -test). In general, the behavior of our model on this dataset is remarkably similar to that on the test subset of Corel5K as shown in Fig. 10(c) and (d) and Siamese-CE always performs better than CE. Unfortunately, all other models perform poorly; most of models yield the performance roughly identical to the Random models', as seen in Fig. 12(c) and (d). In general, the performance of our model is consistent in priming and extended priming on this dataset. Hence, the same conclusion on the priming can be drawn on the extended priming.

Fig. 12(e) and (f) show the extended priming results on SUNDatabase. Once again, our model performs statistically significant ( $p$ -value  $< 0.01$ , Student's  $t$ -test) better than all other models; Siamese-CE wins over the runner-up 22% in MAP and 13% in AUC and, in particular, the significantly better performance at all 11 recall levels. It is also observed that all the models perform on this dataset very similarly to those on LabelMe.

In summary, the CCT experimental results demonstrate that the semantics learnt by our model trained on a dataset may be transferable to other collections if different annotators have a high agreement on the intended meanings of terms and there are sufficient training documents reflecting various concepts and intended term-use patterns. Without meeting the requirement, all the models encounter the same problem in generalization of learnt semantics cross corpora.



**Fig. 12.** Extended priming accuracy of different models in terms of  $P@K$ , MAP, Precision vs. Recall and AUC in the CCT experiments. (a–b) MSD. (c–d) LabelMe. (e–f) SUNDatabase.

### 6.5. Incomplete local context results

In the incomplete local context experiments, we randomly remove a number of terms from an evaluation document to synthesize an incomplete local context with two missing rates, up to 10% and between 10% and 30%, as discussed in Section 6.2.3. The training subsets in MagTag5K, CAL500 and Corel5K are used in this experimental setting and we report the experimental results in terms of the priming and the extended priming protocols. It is also worth clarifying that the CRBM is generally ineligible as its local context is the ID of a query document and hence cannot be distorted. Nevertheless, we use the CRBM only in the extended priming protocol although its local context is not distorted.

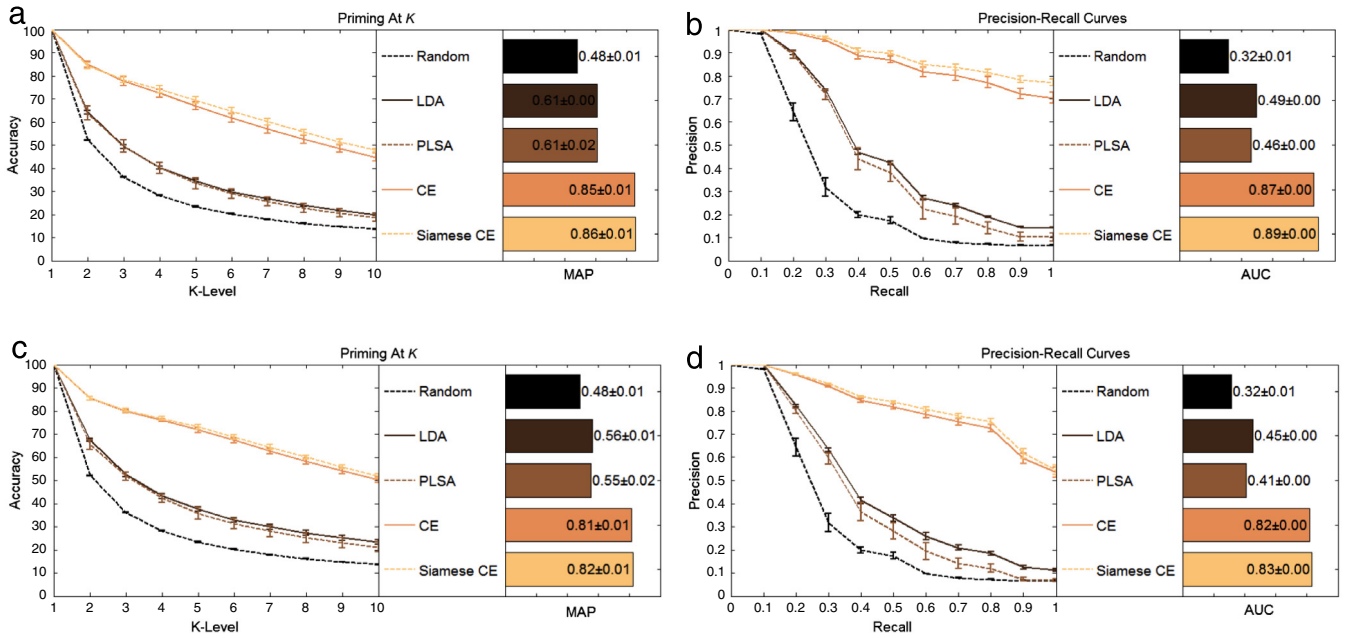
#### 6.5.1. Priming results

Fig. 13 illustrates the priming results of four different models at two missing rates on MagTag5K in terms of four evaluation criteria defined in Section 6.1.3. As expected, it is observed from Fig. 13 that the use of incomplete local context results in the degraded performance for our model due to information loss. In comparison to the results with the complete local context, the performance of Siamese-CE shown in Fig. 13 is lower than those shown in Fig. 5(a) and (b) by 0% and 4% in MAP as well as 1% and 7% in AUC at two missing rates, respectively. In particular, the incomplete local context generally causes the performance at high recall levels to be degraded more than that at low recall levels. In contrast, two PTMs, LDA and PLSA, show robust performance in resisting noisy data as their performance on the incomplete

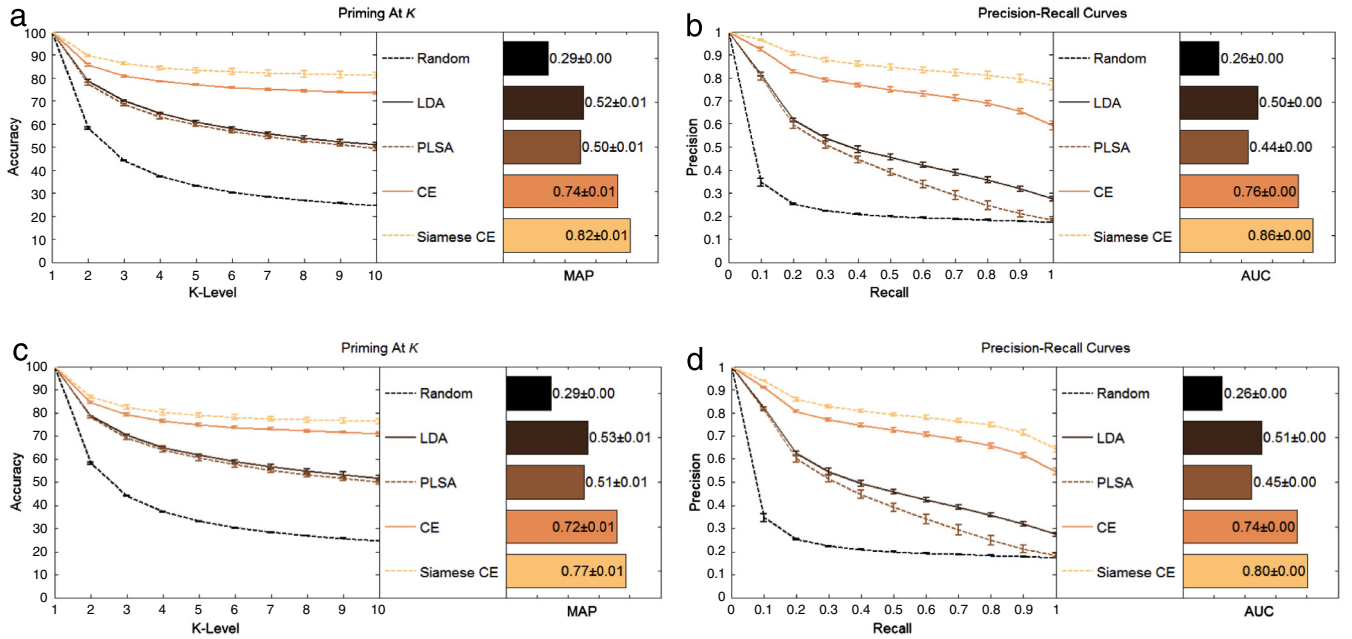
documents is comparable to that of the corresponding complete version. Nevertheless, our model still outperforms other models with the statistical significance ( $p$ -value  $< 0.01$ , Student's  $t$ -test) as seen in Fig. 13.

Fig. 14 shows the priming results of four different models at two missing rates on CAL500 in terms of four evaluation criteria. On this dataset, all the models exhibit almost the same behavior as they work on MagTag5K in the presence of incomplete local context. In comparison to the results with the complete local context, the performance of Siamese-CE on this dataset is reduced by 2% and 7% in MAP as well as 2% and 8% in AUC at two missing rates, respectively. Unlike the behavior on MagTag5K, however, it is observed from Fig. 15 that the performance of Siamese-CE at high recall levels does not decrease sharply. Although LDA and PLSA yield the robust performance, our model still generates the statistically significant ( $p$ -value  $< 0.01$ , Student's  $t$ -test) better performance than all other models including LDA and PLSA in this experimental setting, as is evident in Fig. 15.

Fig. 15 shows the priming results of four different models at two missing rates on Corel5K in terms of four evaluation criteria. It is observed from Fig. 15 that our model exhibits the better robustness in the presence of incomplete local context. On this dataset, the performance of Siamese-CE is only reduced by 0% and 2% in MAP as well as 0% and 4% in AUC at two missing rates, respectively, in comparison to those with complete local context shown in Fig. 7(a) and (b). Recall that on average there are only 3.5 labels in this dataset. By dropping up to 30% terms per document, on average, there are less than two labels per document to form local context.



**Fig. 13.** Priming accuracy of different models on MagTag5K at two missing rates. (a–b) Up to 10%. (c–d) Between 10% and 30%.



**Fig. 14.** Priming accuracy of different models on CAL500 at two missing rates. (a–b) Up to 10%. (c–d) Between 10% and 30%.

Thus, our model leads to favorable results on this dataset. Once again, our model yields the statistically significant ( $p$ -value  $< 0.01$ , Student's  $t$ -test) better performance than other models on Corel5K, as is evident in Fig. 15.

#### 6.5.2. Extended priming results

Fig. 16 illustrates the extended priming results of nine different models at two missing rates on MagTag5K in terms of four evaluation criteria defined in Section 6.1.3. It is observed from Figs. 13 and 16 that the behavior of our model generally remains consistent in two different priming protocols; the performance gradually decreases as the missing rate increases and a higher missing rate causes Siamese-CE to have a sharper performance reduction at high recall levels. With the incomplete local context, the performance of Siamese-CE is reduced by 4% and 1% in MAP

as well as 2% and 6% in AUC at two missing rates, respectively, in comparison to those with complete local context. It is also observed from Fig. 16 that unlike our model, other models perform irregularly, e.g., the performance at a higher missing rate is even better than that at a lower missing rate. Overall, their performance is still significantly inferior to ours.

Fig. 17 shows the extended priming results of nine different models at two missing rates on CAL500 in terms of four evaluation criteria. We observe that on this dataset, all the models including ours behave similarly in comparison to their behavior on MagTag5K; all other models behave irregularly at different missing rates while the performance of our model gradually decreases as the missing rate increases. With the incomplete local context, the performance of Siamese-CE is reduced by 1% and 4% in MAP as well as 2% and 5% in AUC at two missing rates, respectively, in



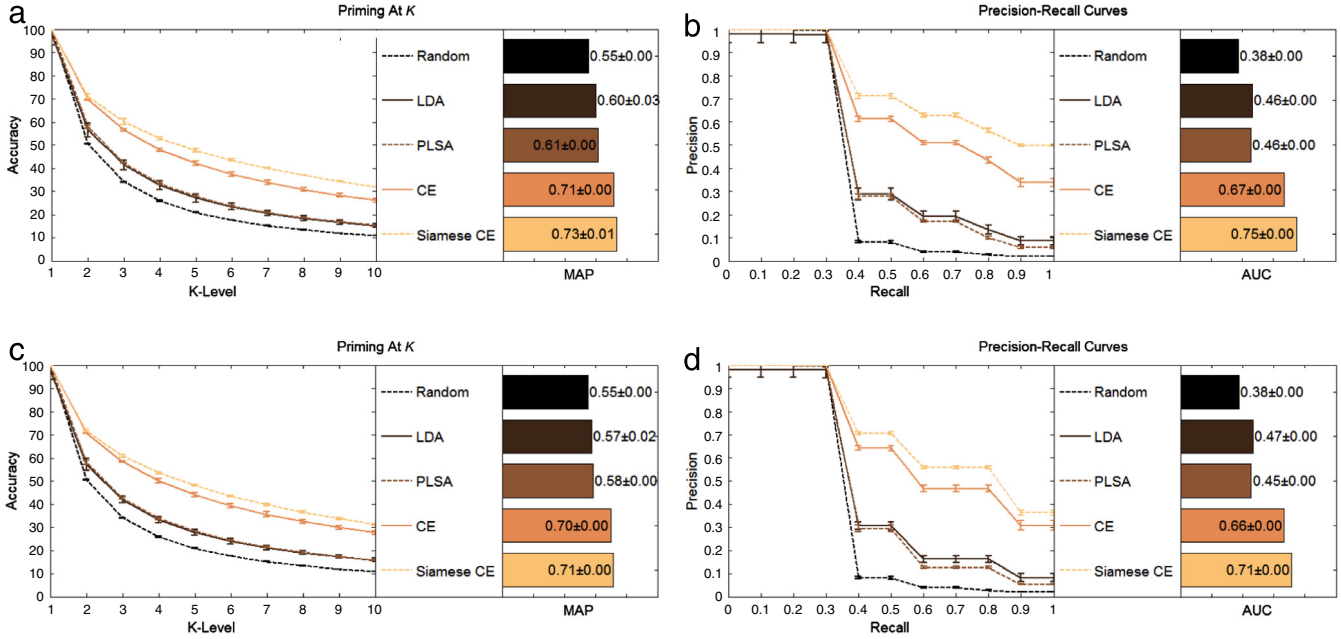


Fig. 15. Priming accuracy of different models on Corel5K at two missing rates. (a–b) Up to 10%. (c–d) Between 10% and 30%.

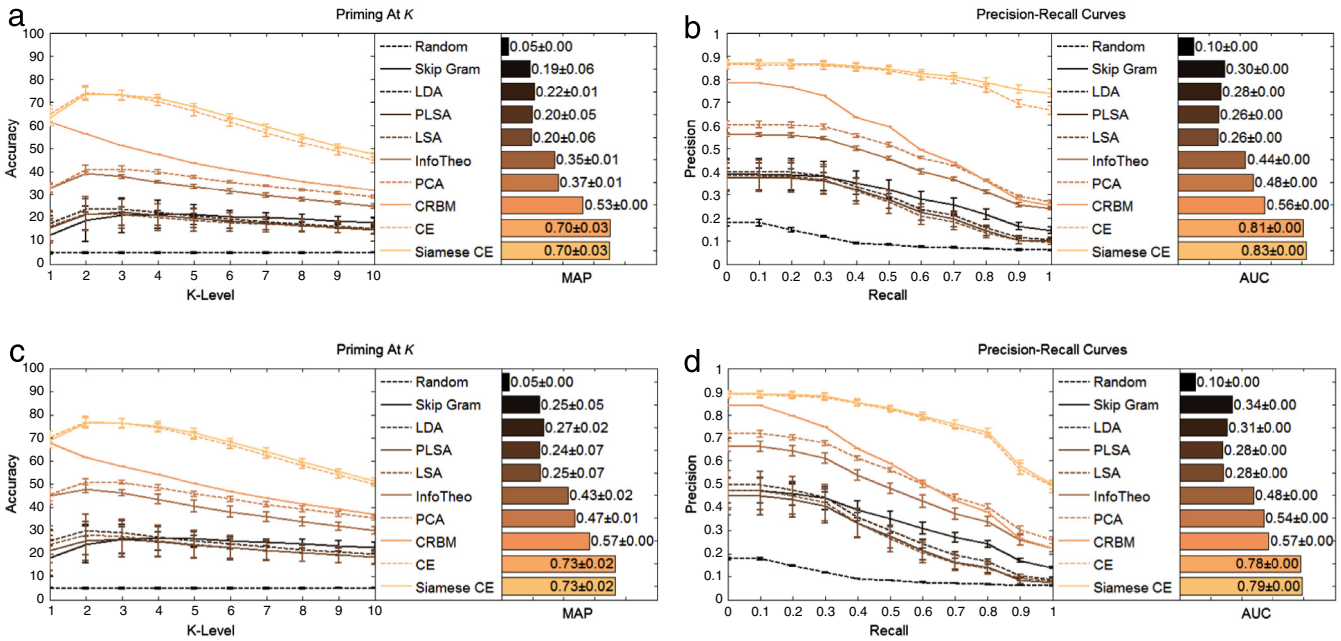


Fig. 16. Extended priming accuracy of different models on MagTag5K at two missing rates. (a–b) Up to 10%. (c–d) Between 10% and 30%.

comparison to those with the complete local context. Thanks to the use of the complete local context, i.e., the document ID, CRBM performs well at all two missing rates although it is still inferior to ours overall. Despite the performance reduction, our model still outperforms all other models on this dataset and, in particular, can prime a few top related terms correctly, as is evident in Fig. 17.

Fig. 18 illustrates the extended priming results of nine different models at two missing rates on Corel5K in terms of four evaluation criteria. Unlike its behavior on two music datasets, our model performs much better; the performance of Siamese-CE is only reduced by 1% in MAP as well as 4% in AUC at two missing rates, respectively, in comparison to those with the complete local context. On this dataset, our model also behaves consistently in two different priming protocols as shown in Figs. 15 and 18. In contrast, all other models have the same behavior as they exhibit

on two music datasets. In general, our model still outperforms other models at all two missing rates in terms of all four evaluation criteria.

In summary, the experimental results in the presence of incomplete local context suggest that our model performs reasonably well despite the performance reduction as expected and generally outperforms all other models on three different datasets in both the priming and the extended priming protocols. In contrast, other models perform irregularly at different missing rates. It is well known that accompany terms in a document may not convey equal amount information. Dropping a term randomly may exert different impacts on its local context. On one hand, it may incur huge information loss and concept change if the term is very informative. On the other hand, it could make little impact if the term is redundant. In general, removing terms randomly from



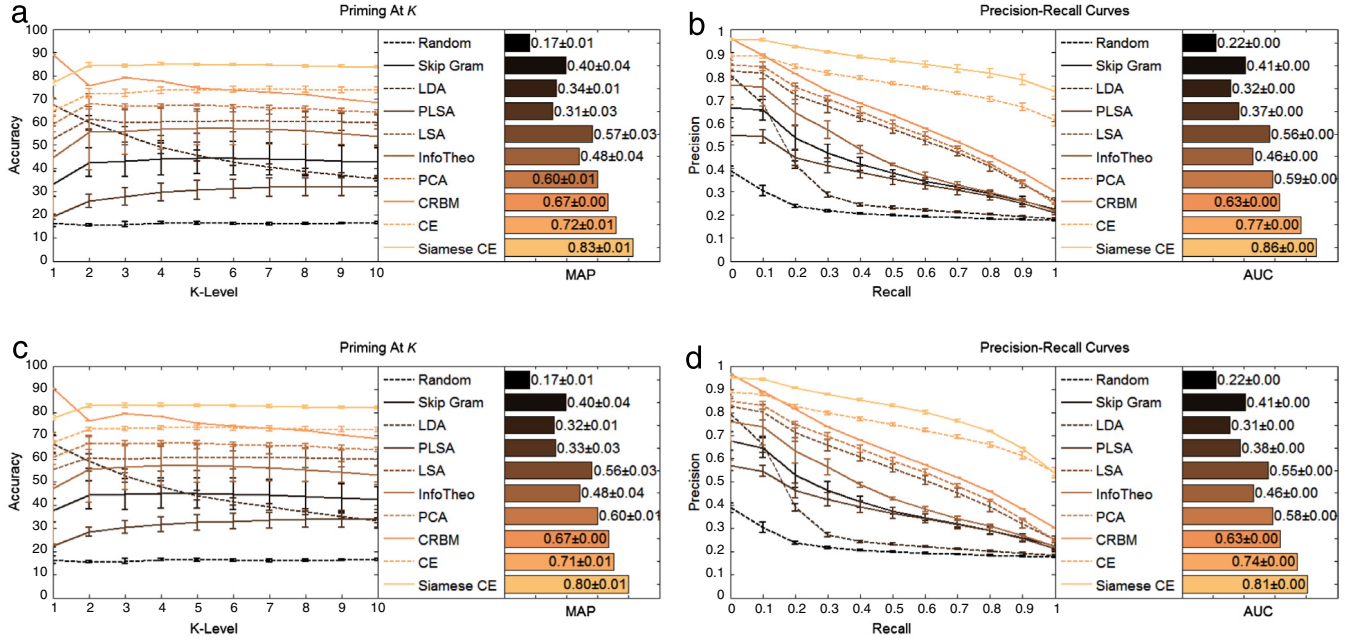


Fig. 17. Extended priming accuracy of different models on CAL500 at two missing rates. (a–b) Up to 10%. (c–d) Between 10% and 30%.

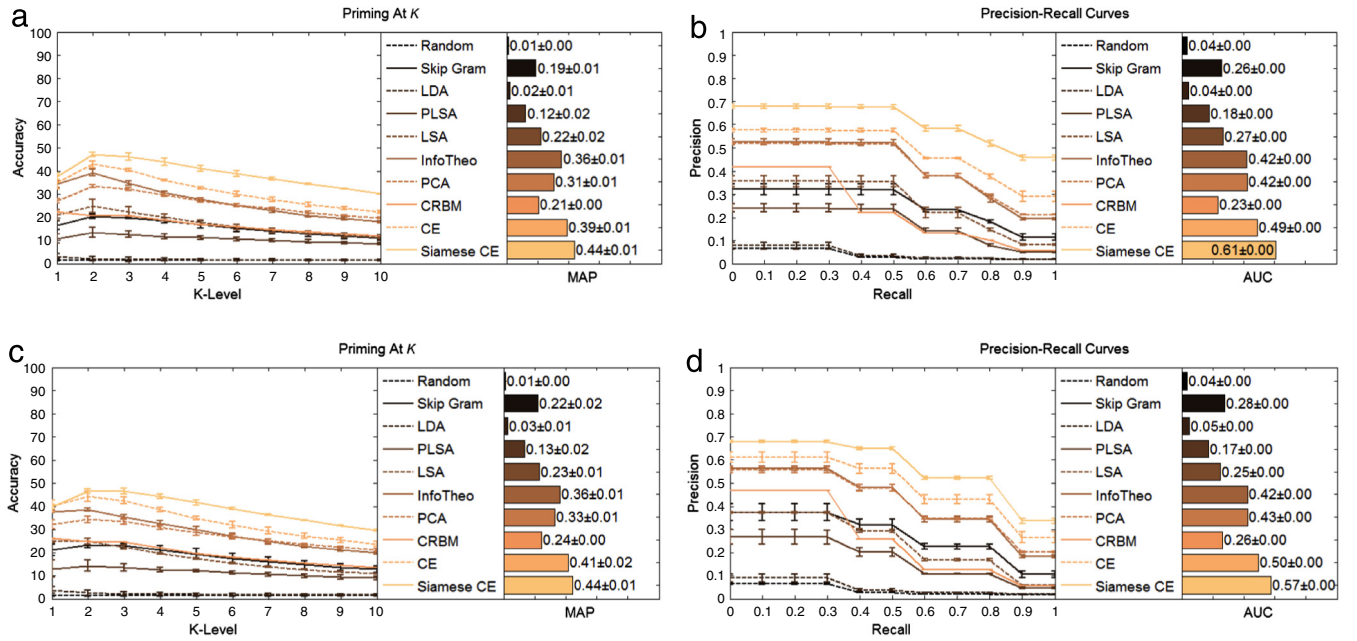


Fig. 18. Extended priming accuracy of different models on Corel5K at two missing rates. (a–b) Up to 10%. (c–d) Between 10% and 30%.

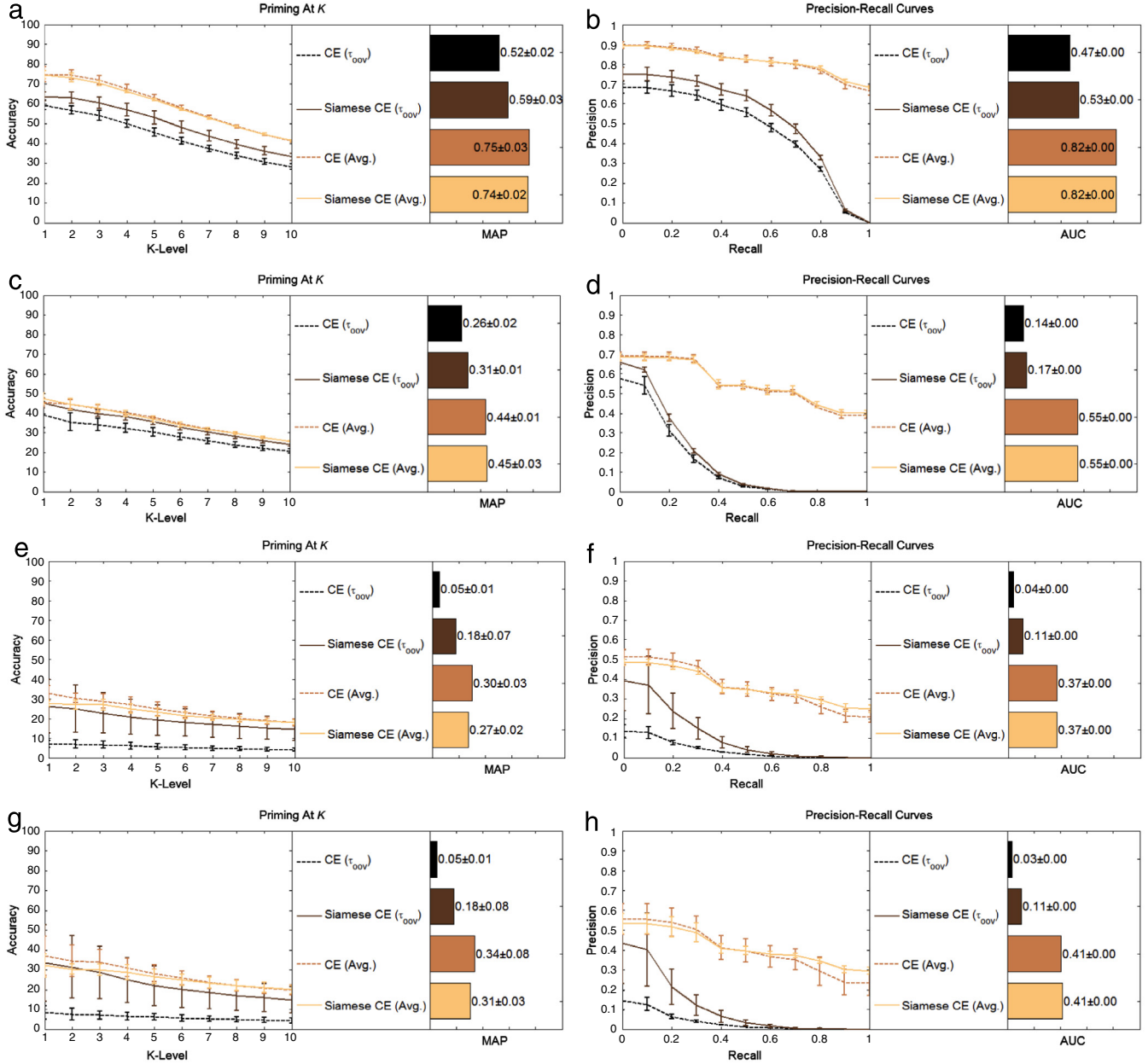
a document may even cause the loss in coherence of co-occurring terms in an incomplete document. Perhaps this setting might be responsible for irregular yet unstable behavior of other models and ours on a number of occasions, e.g., on CAL500.

### 6.6. Out of vocabulary results

With the experimental setting described in Section 6.2.4, we report the OOV experimental results on the reserved OOV set in MagTag5K and the real documents OOV sets in MSD, LabelMe and SUNDatabase in terms of two priming protocols. We use  $\text{CE}(\tau_{\text{OOV}})$  and Siamese- $\text{CE}(\tau_{\text{OOV}})$  to indicate the representations achieved by the feature-based OOV method, i.e., priming related terms using query concept projection  $\text{CE}(\tau_{\text{OOV}}|\delta_{iv})$ , and  $\text{CE}(\text{Avg})$  and Siamese- $\text{CE}(\text{Avg})$  to denote the representations achieved by the

concept-based method, i.e., priming related terms using query concept projection,  $\text{CE}(\mathbf{x}(\tau_{\text{OOV}}, \delta_{iv}))$  (cf. Section 4.4).

Fig. 19 illustrates the priming results of two OOV methods on four different datasets in terms of four evaluation criteria defined in Section 6.1.3. It is evident from Fig. 19 that two proposed OOV methods yield favorable results on different datasets, and  $\text{CE}(\text{Avg})$  and Siamese- $\text{CE}(\text{Avg})$  significantly outperform  $\text{CE}(\tau_{\text{OOV}})$  and Siamese- $\text{CE}(\tau_{\text{OOV}})$  constantly on all four datasets. In particular, the performance of  $\text{CE}(\text{Avg})$  and Siamese- $\text{CE}(\text{Avg})$  on the OOV set is roughly comparable to the performance of CE and Siamese-CE on the test subset of MagTag5K, as is evident in Figs. 5(c), (d), 19(a) and (b). For the feature-based method, Siamese- $\text{CE}(\tau_{\text{OOV}})$  performs considerably better than  $\text{CE}(\tau_{\text{OOV}})$  on all four datasets thanks to the distance learning undertaken by our Siamese architecture. However, the performance of  $\text{CE}(\text{Avg})$  is marginally higher than



**Fig. 19.** Priming accuracy of two OOV methods on different datasets. (a–b) MagTag5K. (c–d) MSD. (e–f) LabelMe. (g–h) SUNDatabase.

that of Siamese-CE(Avg) on all four datasets. It implies that a subnetwork trained for the prediction has embedded the related concepts conveyed in a document reasonably well so that their centroid in the CE space can be used to approximately embed a related OOV term that shares the same local context. On the other hand, the distance learning undertaken by the Siamese architecture is dedicated to the accurate concept embedding concerning only in-vocabulary terms based on training data. Based on the experimental results shown in Fig. 19, we could make better use of CE and Siamese-CE; we use CE to represent OOV terms while Siamese-CE is used only for in-vocabulary terms. Here, we emphasize that our proposed OOV methods directly generate the same CE representation for an OOV term as that of an in-vocabulary term. For any applications that employ our learnt semantics, there is no additional processing required for OOV terms. In other words, both in-vocabulary and OOV terms can be represented uniformly in the CE space.

In summary, the experimental results reported in Sections 6.3–6.6 provide solid evidence to support our problem formulation and

the proposed solution as our approach significantly outperforms other state-of-the-art semantic learning methods. In semantic priming, our approach exhibits its strength in capturing accurate semantics from training corpora and, more importantly, the capability of generalizing the semantics to unseen documents in different situations, noisy documents (resulting in incomplete local context) and documents containing OOV terms. Thus, we believe that our approach is ready for different MMIR applications.

## 7. Discussion

In this section, we discuss several issues arising from our work and relate our approach to previous work in learning semantics from descriptive terms.

To learn semantics from descriptive terms, most of existing techniques often undergo a preprocessing stage by filtering out rarely used terms from those documents concerned (Law et al., 2010; Mandel et al., 2011). Our observations suggest that some

rarely used terms may play a critical role in local context to facilitate understanding the accurate meaning of a specific term. Therefore, our approach always uses all the natural documents without removing any rarely used terms in our training and test. On the other hand, we realize that the frequently used terms may convey commonly used semantics and hence need to be handled differently from rarely used terms. Technically, this could be done in a semantic representation space by using the frequency of terms to normalize the distance between different terms. As this is an issue relating to a specific application, this potential solution needs to be investigated when our approach is applied to a specific task.

In our approach, a novel Siamese architecture and its two-stage learning procedure are proposed especially for learning the concept embedding (CE) from co-occurring terms. As a result, two **CE** representations, CE and Siamese-CE, could be obtained in the first and the second learning stages, respectively; i.e., CE learnt from the prediction task and Siamese-CE generated by working on the distance learning and the prediction task simultaneously. From the experimental results in semantic priming, we observe that Siamese-CE outperforms CE whenever there are sufficient training data reflecting different concepts conveyed in descriptive terms and various intended term usage patterns, while CE may perform slightly better than Siamese-CE for unseen documents in different corpora if the training data do not comply with the aforementioned conditions, e.g., training on MagTag5K, a small music annotation dataset, and test on MSD, a huge and highly diversified music annotation dataset. As the distance learning for Siamese-CE is dedicated to accurate concept embedding based on information carried in training documents, the lack of sufficient information on various concepts and intended usage patterns in training data is responsible for this problem. In contrast, CE does not involve in the refinement via distance learning and hence does not overfit those limited concepts and intended usage patterns in MagTag5K, which leads to the better generalization on MSD. In general, we argue that our Siamese architecture should be applied in generating the **CE** representations if it can be trained on a highly informative dataset, e.g., MSD in the music domain, and the computational efficiency issues arising from our raw representations can be addressed properly as discussed next.

Apparently, our proposed approach relies on the Siamese architecture of deep neural networks to learn complex contextualized semantics from descriptive terms. In general, training a deep neural network involves the non-convex optimization and tedious hyper-parameter tuning. Our proposed approach is inevitably subject to this limitation. Furthermore, we employ the *tfidf* representation to characterize a term (used as a part of input to the deep neural network) and the BoW to represent the coherent terms in a training document (as “target” to learn the prediction). Both representations have the same number of features, equal to the size of the term vocabulary in a training dataset. For a large term vocabulary in a dataset, e.g., Million Song Dataset (MSD), our approach suffers from a heavy computational burden, which prohibits us from training our Siamese architecture on a dataset like MSD with our current computational facility. In general, a parsimonious representation of a large word vocabulary is demanded by various natural language processing tasks and has been studied previously. The potential solutions include applying a dimensionality reduction technique, e.g., PCA or compressed sensing (Hsu, Kakade, Langford, & Zhang, 2009), to the representation and transforming the high-dimensional binary BoW representation into a low-dimensional continuous yet compressed representation (Hsu et al., 2009). While the potential solutions still need to be investigated, we anticipate that such techniques would effectively reduce the computational burden in our approach.

For our proposed Siamese architecture, there are several salient characteristics that distinguish our architecture from most of

existing Siamese architectures. First of all, most of existing Siamese architectures are developed to learn a distance metric only in the representation space (Bordes, Weston, Collobert, & Bengio, 2011; Bromley et al., 1993; Chopra, Hadsell, & LeCun, 2005). Unlike those architectures that learn a single task, ours not only learns a distance metric in the CE space but also simultaneously establishes a predictor that infers the coherent terms from an instance consisting of the raw representations of a focused term and its local context. Next, the existing Siamese architectures are generally trained via supervised learning. In contrast, ours is trained in two stages via unsupervised learning. Finally, ours is also different from those regularized Siamese architectures (Chen & Salman, 2011; Salakhutdinov & Hinton, 2007). Such regularized variants employ an auto-encoder as its subnetwork in order to minimize information loss so as to achieve better generalization, while ours learns two relevant yet different tasks, prediction and distance metric learning, simultaneously. In addition, those regularized Siamese architectures are still trained via supervised learning. Apart from those salient characteristics, we believe that our proposed Siamese architecture and learning algorithms can be easily extended to other types of contextualized semantic learning from descriptive terms by means of alternative context information instead of our used local context, i.e., co-occurring terms in an annotation document, defined in this paper.

In this paper, we have formulated a contextualized semantic learning task from collections of textual descriptive terms independent of any specific MMIR application tasks. We believe that a solution to this problem would facilitate bridging the semantic gap between media content and relevant high-level concepts. Our work presented in this paper is different from the previous studies phrased with term “contextual” in this area. For example, the “contextual object recognition” (Rasiwasia & Vasconcelos, 2012) actually refers to a method of exploiting the relatedness of object labels achieved from training an object recognizer (a multi-class classifier) to improve the performance by using such information as context to train another classifier. That “context” achieved directly from media content for a specific application task is by no means relevant to our contextualized semantic learning task and the proposed solution. In addition, the “contextual tag inference” (Mandel et al., 2011) is an approach that exploits descriptive terms in order to produce a smoothed representation for documents with CRBM. The smoothed representation is a document-level summary of the document-term relatedness to improve the term-based auto-annotation performance via providing smoothed target labels instead of binary ones. This smoothed representation acts as a novel document-level representation but does not capture the term-to-term relatedness explicitly. In other words, this method does not provide an explicit continuous embedding representation for a term. Technically, this approach is subject to limitation in generalization as the learnt representation is merely applicable to the training documents due to the context characterized by the document ID. Although this limitation might be overcome by using some alternative contextual information, this method is still not a legitimate solution to our formulated problem due to the lack of continuous embedding representation. In nature, the work closest to ours is the “Association Rules” (Agrawal, Imieliński, & Swami, 1993) that lead to contextualized semantic representations on a conceptual level (Yang, Huang, Shen, & Zhou, 2010). Unlike the local context used in our approach, different rules may apply different contextual information for concept modeling, which could result in inconsistency due to the intricate contextual information. Furthermore, the mined rules can provide binary concept-to-concept relatedness only, which confines itself to a limited range of applications.

In general, the solution presented in this paper leads to multiple continuous **CE** representations for a descriptive term depending on



the local context. In essence, one **CE** representation of a term tends to accurately model a concept intended by annotators. Moreover, the **CE** representation space and the learnt semantic distance metric allow similar concepts to associate with each other and make different concepts readily distinguished. As a result, our **CE** representation scheme significantly distinguishes from other semantic representations learnt from descriptive terms. Without taking contextual information into account, the learnt semantic representation reflects only the global term-to-term relatedness and hence each term has a unique representation (Deerwester et al., 1990; Markines et al., 2009). In the existing work in addressing contextualized semantics, most of those methods, e.g., smoothing (Mandel et al., 2011) and probabilistic topic models, LDA (Blei et al., 2003) and PLSA (Hofmann, 1999), offer only a document-level representation but do not address the contextualized term-to-term relatedness issue directly. It is well known that several documents may together specify a single concept, while one document may convey multiple concepts. Therefore, a concept-level representation is always required even though an application works on the level of documents relatedness. Here, we emphasize that for a document of  $m$  terms, we employ all  $m$  concept-level representations arising from this document collectively to form a document-level representation and the learnt semantic distance metric by our Siamese architecture can easily adapt to measuring the document-to-document relatedness as demonstrated in the extended semantic priming.

Finally, it is worth mentioning that there are alternative methods for solving our formulated problem. Most of those methods fall into the ontology area and rely on human expertise such as “tag ontologies” (Kim et al., 2008; Wang et al., 2010) and “property lists” semantics (Sun, Kim, Kohli, & Savarese, 2013). On the one hand, experts’ specialist knowledge regarding terms and their inter-relatedness is harvested so that a system can use such semantics to perform human-like tasks. On the other hand, such acquisition of semantics has to be handcrafted and time-consuming, which is laborious and hence incurs a huge cost. Moreover, the acquired semantics may result in experts’ bias, and the subjective opinion differences may even cause conflicting semantics. In contrast, our approach presented in this paper lends clear evidence to favor learning semantics from descriptive terms as our approach is much less costly and can automatically capture concepts underlying terms in context by following trends of the crowd in meaning.

## 8. Conclusion

We have presented an approach to acquiring contextualized semantics from co-occurring descriptive terms. In our approach, we have formulated the problem as learning a contextualized term-based semantic representation via concept embedding in the representation space. As a result, we have proposed a solution by developing a novel Siamese architecture of deep neural networks and a two-stage learning algorithm. We have also addressed the OOV issues in our solution. By means of visualization, we have demonstrated that our approach can capture domain-specific and transferable contextualized semantics conveyed in co-occurring terms. Moreover, we have applied our approach to semantic priming, a benchmark information retrieval task. We have conducted a thorough evaluation via a comparative study with different settings. Experimental results suggest that our approach outperforms a number of state-of-the-art approaches and the effectiveness of our proposed OOV methods in this benchmark task.

While our proposed Siamese architecture and learning algorithms provide a solution to the formulated problem, there are still several issues to be tackled including the computational efficiency in using a training corpus of a large term vocabulary; exploring alternative contextual information sources and modeling

techniques; and extension of our Siamese architecture to learning other types of contextualized semantics other than the local context defined in this paper. In our ongoing work, we shall be dealing with the aforementioned issues and applying the learnt contextualized semantic representations to a number of real MMIR applications, e.g., auto-annotation of multimedia content, term/media content recommendation and query expansion, multimedia retrieval with textual queries as well as zero-shot learning in various multimedia classification tasks. We anticipate that the formulated problem and our solution presented in this paper would pave a new way towards bridging the semantic gap.

## Acknowledgments

The authors would like to thank the action editor and anonymous reviewers for their comments that improve the presentation of this manuscript.

## Appendix

In this appendix, we derive the learning algorithms used to train our proposed architecture. To minimize the loss functions defined for the prediction and the distance metric learning described in Section 4 of the main text, we use the stochastic back propagation (SBP) algorithm for parameter estimation. To establish a deep subnetwork, the pre-training is carried out in a greedy layer-wise fashion with each layer’s weights obtained via training a sparse autoencoder with a Quasi-Newton method as described in Appendix A.1. In Appendices A.2 and A.3, we present the derivation of gradients of loss functions with respect to relevant parameters used to train a subnetwork for the prediction and to train the Siamese architecture for the distance metric learning, respectively. Finally, we summarize the SBP algorithm that can be used in training one subnetwork for the prediction and the Siamese architecture for the distance learning.

### A.1. Sparse auto-encoder learning

A sparse autoencoder can be used to initialize weights of a deep neural network by reconstructing the input via a single hidden and preferably sparse layer (Ranzato, Boureau, & LeCun, 2007). In our experiments, the sparse autoencoder was trained using batch training using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method, a variant of Quasi-Newton method in the popular implementation of minFunc (Schmidt, 2005).

Let  $\mathbf{x}$  be an input vector. The hidden layer’s activations are

$$\mathbf{z}_1(\mathbf{x}) = \mathbf{f}(W_1\mathbf{x} + \mathbf{b}_1),$$

and the corresponding output layer’s activations are

$$\tilde{\mathbf{x}}(\mathbf{x}) = \mathbf{f}(W_2\mathbf{z}_1(\mathbf{x}) + \mathbf{b}_2),$$

where  $W_1$ ,  $\mathbf{b}_1$ ,  $W_2$  and  $\mathbf{b}_2$  are encoding weights and biases and decoding weights and biases, respectively.  $\mathbf{f}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$  is the hyperbolic tangent function used in our experiments.

Encouraging sparsity is carried out via a regularizer to the cost which consists of penalizing the magnitude of the hidden layer’s output regardless of the sign:

$$\mathcal{R} = \sum_{q=1}^Q \sqrt{(\mathbf{z}_1(\mathbf{x})[q])^2 + \epsilon},$$

where  $Q$  is the number of units in the hidden layer.

The objective of the training is minimizing the following loss averaged over all the examples:

$$\mathcal{L}_A(X; \Theta) = \sum_{k=1}^K \left\| \tilde{\mathbf{x}}(\mathbf{x}_k) - \mathbf{x}_k \right\|_2^2 + \alpha \sum_{k=1}^K \left( \sum_{q=1}^Q \sqrt{(\mathbf{z}_1(\mathbf{x})[q])^2 + \epsilon} \right).$$

Hence, we achieve  $\frac{\partial \mathcal{L}_A(X; \Theta)}{\partial \tilde{\mathbf{x}}(\mathbf{x})} = 2 \left\| \tilde{\mathbf{x}}(\mathbf{x}) - \mathbf{x} \right\|_1$  and  $\frac{\partial \mathcal{R}}{\partial \mathbf{z}_1(\mathbf{x})} = \frac{\mathbf{z}_1(\mathbf{x})}{\sqrt{(\mathbf{z}_1(\mathbf{x}))^2 + \epsilon}}$ .

Let  $\nabla \mathbf{f}(\mathbf{x})$  be the gradient of the hyperbolic function given input  $\mathbf{x}$ . We have

$$\nabla \mathbf{f}(\mathbf{x}) = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \nabla \tanh(\mathbf{x}) = 1 - (\tanh(\mathbf{x}))^2.$$

Given a training dataset  $X$  of  $K$  examples, we apply the chain rule in order to obtain the derivatives with respect to a specific parameter as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial W_2} &= \sum_{k=1}^K \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)} \cdot \frac{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)}{\partial W_2} \\ &= 2 \sum_{k=1}^K ((\tilde{\mathbf{x}}(\mathbf{x}_k) - \mathbf{x}_k) \cdot \nabla \mathbf{f}(W_2 \mathbf{z}_1(\mathbf{x}_k) + \mathbf{b}_2) \cdot \mathbf{z}_1(\mathbf{x}_k)) \\ \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial b_2} &= \sum_{k=1}^K \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)} \cdot \frac{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)}{\partial b_2} \\ &= 2 \sum_{k=1}^K ((\tilde{\mathbf{x}}(\mathbf{x}_k) - \mathbf{x}_k) \cdot \nabla \mathbf{f}(W_2 \mathbf{z}_1(\mathbf{x}_k) + \mathbf{b}_2)) \\ \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial W_1} &= \sum_{k=1}^K \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)} \cdot \frac{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)}{\partial W_1} \cdot \frac{\partial \mathbf{z}_1(\mathbf{x}_k)}{\partial W_1} \\ &\quad + \alpha \frac{\partial \mathcal{R}}{\partial \mathbf{z}_1(\mathbf{x}_k)} \cdot \frac{\partial \mathbf{z}_1(\mathbf{x}_k)}{\partial W_1} \\ &= 2 \sum_{k=1}^K ((\tilde{\mathbf{x}}(\mathbf{x}_k) - \mathbf{x}_k) \cdot \nabla \mathbf{f}(W_2 \mathbf{z}_1(\mathbf{x}_k) + \mathbf{b}_2) \cdot W_2 \cdot \nabla \mathbf{f}(W_1 \mathbf{x}_k + \mathbf{b}_1) \cdot \mathbf{x}_k) \\ &\quad + \alpha \sum_{k=1}^K \left( \frac{\mathbf{z}_1(\mathbf{x}_k)}{\sqrt{(\mathbf{z}_1(\mathbf{x}_k))^2 + \epsilon}} \cdot \nabla \mathbf{f}(W_1 \mathbf{x}_k + \mathbf{b}_1) \cdot \mathbf{x}_k \right) \\ \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial b_1} &= \sum_{k=1}^K \frac{\partial \mathcal{L}_A(X; \Theta)}{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)} \cdot \frac{\partial \tilde{\mathbf{x}}(\mathbf{x}_k)}{\partial b_1} \cdot \frac{\partial \mathbf{z}_1(\mathbf{x}_k)}{\partial b_1} \\ &\quad + \alpha \frac{\partial \mathcal{R}}{\partial \mathbf{z}_1(\mathbf{x}_k)} \cdot \frac{\partial \mathbf{z}_1(\mathbf{x}_k)}{\partial b_1} \\ &= 2 \sum_{k=1}^K ((\tilde{\mathbf{x}}(\mathbf{x}_k) - \mathbf{x}_k) \cdot \nabla \mathbf{f}(W_2 \mathbf{z}_1(\mathbf{x}_k) + \mathbf{b}_2) \cdot W_2 \cdot \nabla \mathbf{f}(W_1 \mathbf{x}_k + \mathbf{b}_1)) \\ &\quad + \alpha \sum_{k=1}^K \left( \frac{\mathbf{z}_1(\mathbf{x}_k)}{\sqrt{(\mathbf{z}_1(\mathbf{x}_k))^2 + \epsilon}} \cdot \nabla \mathbf{f}(W_1 \mathbf{x}_k + \mathbf{b}_1) \right). \end{aligned} \quad (\text{A.1})$$

The sparse auto-encoder is employed to initialize a subnetwork recursively where each layer is trained based on the output of its previous layer until a specified number of layers are achieved.

## A.2. Subnetwork learning for prediction

As defined in Eq. (4) of the main text, the prediction loss is

$$\mathcal{L}_P(X; \Theta) = -\frac{1}{2K |\Gamma|} \sum_{k=1}^K \mathcal{L}_P(\mathbf{x}_k; \Theta),$$

where  $\mathcal{L}_P(\mathbf{x}_k; \Theta) = \sum_{i=1}^{|\Gamma|} (\kappa_k (1 + y_k[i]) \log(1 + \hat{y}_k[i]) + (1 - \kappa_k) (1 - y_k[i]) \log(1 - \hat{y}_k[i]))$ .

Here,  $\hat{y}_k[i]$  and  $y_k[i]$  represent the prediction and the true label related to term  $i$  in example  $k$ , respectively. By applying the chain rule, we have

$$\frac{\partial \mathcal{L}_P(X; \Theta)}{\partial \mathbf{x}} = \frac{1}{K} \sum_{k=1}^K \left( \frac{\partial \mathcal{L}_P(\mathbf{x}_k; \Theta)}{\partial \hat{\mathbf{y}}_k} \cdot \frac{\partial \hat{\mathbf{y}}_k}{\mathbf{x}} \right),$$

where  $\hat{\mathbf{y}}_k$  is the output vector of prediction, a collective notation of all  $\hat{y}_k[i]$ , also operator is the element wise multiplication. We have

$$\begin{aligned} \frac{\partial \mathcal{L}_P(\mathbf{x}_k; \Theta)}{\partial \hat{\mathbf{y}}_k} &= \frac{-1}{2 |\Gamma|} \sum_{i=1}^{|\Gamma|} \left( \kappa_k \frac{1 + y_k[i]}{1 + \hat{y}_k[i]} \cdot \frac{\partial \hat{y}_k[i]}{\partial \hat{\mathbf{y}}_k} \right. \\ &\quad \left. - (1 - \kappa_k) \frac{1 - y_k[i]}{1 - \hat{y}_k[i]} \cdot \frac{\partial \hat{y}_k[i]}{\partial \hat{\mathbf{y}}_k} \right) \\ &= \frac{-1}{2 |\Gamma|} \left( \kappa_k \frac{1 + \mathbf{y}_k}{1 + \hat{\mathbf{y}}_k} - (1 - \kappa_k) \frac{1 - \mathbf{y}_k}{1 - \hat{\mathbf{y}}_k} \right), \end{aligned}$$

where  $\frac{1+\mathbf{y}}{1+\hat{\mathbf{y}}}$  and  $\frac{1-\mathbf{y}}{1-\hat{\mathbf{y}}}$  are the collective notations of  $\frac{1+y_k[i]}{1+\hat{y}_k[i]}$  and  $\frac{1-y_k[i]}{1-\hat{y}_k[i]}$  with the element-wise division.

Let  $\frac{\partial \mathcal{L}_P(X; \Theta)}{\partial \hat{\mathbf{y}}}$  be the matrix formed by stacking all training  $\frac{\partial \mathcal{L}_P(\mathbf{x}_k; \Theta)}{\partial \hat{\mathbf{y}}_k}$  and  $\mathbf{z}_{H-1}(X)$  be the matrix formed by stacking all  $\mathbf{z}_{H-1}(\mathbf{x}_k)$ . Back propagation starts at the top layer with the partial of the cost on the last layer's parameters (weights and biases),  $\frac{\partial \mathcal{L}_P(X; \Theta)}{\partial W_H}$  and  $\frac{\partial \mathcal{L}_P(X; \Theta)}{\partial \mathbf{b}_H}$  given by

$$\begin{aligned} \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial W_H} &= \frac{1}{K} \sum_{k=1}^K \left( \frac{\partial \mathcal{L}_P(\mathbf{x}_k; \Theta)}{\partial \hat{\mathbf{y}}_k} \cdot \nabla \mathbf{f}(W_H \mathbf{z}_{H-1}(\mathbf{x}_k) + \mathbf{b}_H) \right. \\ &\quad \left. + \mathbf{b}_H \right) \cdot \frac{\partial (W_H \times \mathbf{z}_{H-1}(\mathbf{x}_k) + \mathbf{b}_H)}{W_H} \\ &= \frac{1}{K} \left( \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial \hat{\mathbf{y}}} \cdot \nabla \mathbf{f}(W_H \mathbf{z}_{H-1}(X) + \mathbf{b}_H) \right) * (\mathbf{z}_{H-1}(X))^T, \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial \mathbf{b}_H} &= \frac{1}{K} \sum_{k=1}^K \left( \frac{\partial \mathcal{L}_P(\mathbf{x}_k; \Theta)}{\partial \hat{\mathbf{y}}_k} \cdot \nabla \mathbf{f}(W_H \mathbf{z}_{H-1}(\mathbf{x}_k) + \mathbf{b}_H) \right) \\ &= \frac{1}{K} \left( \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial \hat{\mathbf{y}}} \cdot \nabla \mathbf{f}(W_H \mathbf{z}_{H-1}(X) + \mathbf{b}_H) \right), \end{aligned} \quad (\text{A.2})$$

where  $*$  is the matrix multiplication.

Derivatives with respect to all the parameters,  $W_h$  and  $\mathbf{b}_h$ , of hidden layer  $h = (H - 1, \dots, 1)$  are obtained by the successive use of the chain rule for error back-propagation:

$$\begin{aligned} \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial W_h} &= \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial W_h} \cdot \frac{\partial (W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)}{\partial W_h}, \\ &= \underbrace{\frac{\partial \mathcal{L}_P(X; \Theta)}{\partial (W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)}}_{\zeta_h} \cdot \frac{\partial (W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)}{\partial W_h}, \\ \zeta_h &= \zeta_{h+1} \cdot \frac{\partial (W_{h+1} \mathbf{z}_h(X) + \mathbf{b}_{h+1})}{\partial \mathbf{z}_h(X)} \cdot \frac{\partial \mathbf{z}_h(X)}{\partial (W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)} \\ &= W_{h+1} * (\zeta_{h+1} \cdot \nabla \mathbf{f}(W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)), \end{aligned}$$

$$\begin{aligned} \frac{\partial (W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)}{\partial W_h} &= \mathbf{z}_{h-1}(X), \\ \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial \mathbf{b}_h} &= \frac{\partial \mathcal{L}_P(X; \Theta)}{\partial (W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)} \cdot \frac{\partial (W_h \mathbf{z}_{h-1}(X) + \mathbf{b}_h)}{\partial \mathbf{b}_h} = \zeta_h. \end{aligned} \quad (\text{A.3})$$

### A.3. Siamese architecture learning

As defined in Eq. (7) of the main text, the Siamese loss is

$$\begin{aligned} \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta) &= \frac{1}{K} \sum_{k=1}^K (I_1(\mathbb{E} - \beta(1 - \mathbb{S}))^2 \\ &\quad + I_2\rho(\mathbb{E} - \beta(1 - \mathbb{S}))^2 + I_3(\mathbb{E} - \beta)^2 \mathbb{S}). \end{aligned}$$

Here  $\mathbb{E} = E(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)})$  is the Euclidean distance between the embedding vectors of pairs of input examples and  $\mathbb{S} = e^{-\frac{\lambda}{2} \text{KL}(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)})}$  the target distance is based on contexts similarity following  $\text{KL}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sum_{c=1}^{|\Phi|} ((l^{(1)}[c] - l^{(2)}[c]) \log(\frac{l^{(1)}[c]}{l^{(2)}[c]}))$  where  $|\Phi|$  features in the context representation and  $l^{(i)}[c]$  represents the  $c$ th feature value in the context input provided for subnetwork number  $i$ . Note that  $\beta(1 - \mathbb{S})$  is a constant irrespective of the weights and biases. Moreover, this loss is also unaffected by any weights connected between the CE (i.e., hidden layer  $H - 1$ ) and the prediction layers. Thus,  $\frac{\partial \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta)}{\partial W_H} = 0$  and  $\frac{\partial \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta)}{\partial \mathbf{b}_H} = 0$ .

As two subnetworks always need to be kept identical, all the parameters in each subnetwork are updated by using the averaging derivatives obtained based two subnetworks after each back propagating iteration. As there is no interaction between the two subnetworks apart from the CE layers, we can write

$$\begin{aligned} \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \Theta} &= \frac{1}{2} \left( \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(1)})} \cdot \frac{\partial \mathbf{CE}(\mathbf{x}_k^{(1)})}{\partial \Theta} \right. \\ &\quad \left. + \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(2)})} \cdot \frac{\partial \mathbf{CE}(\mathbf{x}_k^{(2)})}{\partial \Theta} \right). \end{aligned}$$

As the loss is symmetric in terms of the embedding vectors, the derivatives have a uniform form for subnetworks  $i = 1, 2$ :

$$\begin{aligned} \frac{\partial \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta)}{\partial \mathbf{CE}(X^{(i)})} &= \frac{1}{K} \sum_{k=1}^K \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})} \\ \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})} &= 2(I_1(\mathbb{E} - \beta(1 - \mathbb{S})) \\ &\quad + \rho I_2(\mathbb{E} - \beta(1 - \mathbb{S})) \\ &\quad + I_3((\mathbb{E} - \beta)\mathbb{S})) \cdot \frac{\partial E(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)})}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})}. \end{aligned}$$

Focusing on  $\frac{\partial E(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)})}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})}$ , we have

$$\begin{aligned} \frac{\partial E(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)})}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})} &= \frac{\|\mathbf{CE}(\mathbf{x}_k^{(1)}) - \mathbf{CE}(\mathbf{x}_k^{(2)})\|_1}{E(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)})}; E(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}) \\ &= \|\mathbf{CE}(\mathbf{x}_k^{(1)}) - \mathbf{CE}(\mathbf{x}_k^{(2)})\|_2. \end{aligned}$$

Effectively, we can now estimate the partial derivatives for the embedding layer's weights and biases regarding subnetworks  $i = 1, 2$ :

$$\begin{aligned} \frac{\partial \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta)}{\partial W_{H-1}^{(i)}} &= \frac{1}{K} \sum_{k=1}^K \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})} \cdot \frac{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})}{\partial W_{H-1}^{(i)}} \\ &= \frac{1}{K} \sum_{k=1}^K \left( \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})} \cdot \nabla \mathbf{f}(W_{H-1} \mathbf{z}_{H-2}(\mathbf{x}_k^{(i)} \right. \\ &\quad \left. + \mathbf{b}_{H-1})) \right) \cdot \frac{\partial (W_{H-1} \mathbf{z}_{H-2}(\mathbf{x}_k^{(i)} + \mathbf{b}_{H-1}))}{\partial W_{H-1}} \\ &= \frac{1}{K} \left( \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})} \cdot \nabla \mathbf{f}(W_{H-1} \mathbf{z}_{H-2}(\mathbf{x}_k^{(i)} \right. \\ &\quad \left. + \mathbf{b}_{H-1})) \right) * (\mathbf{z}_{H-2}(\mathbf{x}_k^{(i)}))^T, \\ \frac{\partial \mathcal{L}_S(X^{(1)}, X^{(2)}; \Theta)}{\partial \mathbf{b}_{H-1}^{(i)}} &= \frac{1}{K} \sum_{k=1}^K \left( \frac{\partial \mathcal{L}_S(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \Theta)}{\partial \mathbf{CE}(\mathbf{x}_k^{(i)})} \right. \\ &\quad \left. \cdot \nabla \mathbf{f}(W_{H-1} \mathbf{z}_{H-2}(\mathbf{x}_k^{(i)} + \mathbf{b}_{H-1})) \right). \end{aligned} \quad (\text{A.4})$$

The rest of the derivatives are obtained by back propagating the cost in the same fashion as presented in Eq. (A.3).

### A.4. Stochastic gradient descent procedure

Here, we present a generic stochastic gradient descent (SGD) procedure applicable to training a subnetwork for the prediction and the Siamese architecture with the derivatives in Eqs. (A.2)–(A.4). Given a training dataset  $(X, Y)$  where  $X$  is the set of input instances consisting of *tfidf* and context features and  $Y$  is the set of corresponding documents represented in the BoW, the SGD procedure is summarized as follows:

---

**Algorithm: Stochastic Gradient Descent for Loss  $\mathcal{L}(f(X), Y; \Theta)$**

---

**Input:** Initial parameters  $\Theta_0$ , initial learning rate  $\eta_0$  and a stopping threshold  $e_T$ .

**Output:** Optimal parameters  $\Theta_T$

---

```

1:  $t \leftarrow 1$ 
2: Repeat
3:    $\hat{Y}_t \leftarrow f(X; \Theta_t)$ 
4:    $e_t \leftarrow \mathcal{L}(\hat{Y}_t, Y; \Theta_t)$ 
5:    $\nabla \mathcal{L}(t) \leftarrow \text{gradient of } \mathcal{L}(\hat{Y}_t, Y; \Theta_t)$ 
6:    $\Theta_{t+1} \leftarrow \Theta_t - \eta_t \cdot \nabla \mathcal{L}(t)$ 
7:    $\eta_{t+1} \leftarrow \begin{cases} 0.95 * \eta_t & \text{if } (t \bmod 200) = 0 \\ \eta_t & \text{otherwise} \end{cases}$ 
8:    $t \leftarrow t + 1$ 
9: Until  $\frac{e_{t-1}}{e_t} < e_T$ 
10:  $\Theta_T \leftarrow \Theta_t$ 

```

---

It is worth clarifying that the stopping condition in the SGD is generic and applicable to any applications. However, we used a specific stopping condition in our experiments as described in Section 5.2.



## References

- Agrawal, R., Mieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data*. Vol. 22. New York, NY, USA (pp. 207–216).
- Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In B. Schölkopf, J. C. Platt, & T. Hoffman. (Eds.), *Advances in neural information processing systems* 19. Vancouver, Canada (pp. 153–160).
- Bertin-mahieux, T., Ellis, D., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th international society for music information retrieval conference*. Miami, FL, USA (pp. 591–596).
- Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bordes, A., Weston, J., Collobert, R., & Bengio, Y. (2011). Learning structured embeddings of knowledge bases. In *Conference on artificial intelligence*. San Francisco, CA, USA.
- Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G. Orr, & K. Müller (Eds.), *Neural networks, tricks of the trade, reloaded* (pp. 430–445). Berlin, Heidelberg: Springer-Verlag.
- Bottou, L. (2014). From machine learning to machine reasoning: an essay. *Machine Learning*, 94, 133–149.
- Bromley, J., Bentz, J., Bottou, L., Guyon, I., Lecun, Y., Moore, C., et al. (1993). Signature verification using a Siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4), 669–688.
- Chang, S., Eleftheriadis, A., & McClintock, R. (1998). Next-generation content representation, creation, and searching for new-media applications in education. *Proceedings of the IEEE*, 86(5), 884–904.
- Chen, K., & Salman, A. (2011). Learning speaker-specific characteristics with a deep neural architecture. *IEEE Transactions on Neural Networks*, 22(11), 1744–1756.
- Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Conference on computer vision and pattern recognition*. Vol. 1 (pp. 539–546).
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12, 2493–2537.
- Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodology)*, 39(1), 1–38.
- Duygulu, P., Barnard, K., Freitas, N., & Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Conference on computer vision and pattern recognition*. (pp. IV:97–112).
- Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). DeViSE: A deep visual-semantic embedding model. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* 26. Lake Tahoe, NV, USA (pp. 2121–2129).
- Harris, Z. (1954). Distributional structure. *WORD-Journal of the International Linguistic Association*, 10(2–3), 146–162.
- Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 1771–1800.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA (pp. 50–57).
- Hsu, D., Kakade, S., Langford, J., & Zhang, T. (2009). Multi-Label prediction via compressed sensing. In *Advances in neural information processing systems* 22. Vancouver, Canada (pp. 1–14).
- Hueting, M., Monszpart, A., & Mellado, N. (2014). MCGraph: Multi-criterion representation for scene understanding. In *SIGGRAPH Asia 2014 workshop on indoor scene understanding: where graphics meets vision*. Shenzhen, China (pp. 16–24).
- Jaimes, A., Christel, M., Gilles, S., Sarukkai, R., & Ma, W. (2005). Multimedia information retrieval: What is it, and why isn't anyone using it? In *Proceedings of the 7th ACM SIGMM international workshop on multimedia information retrieval*. New York, New York, USA (pp. 3–8).
- Kaminskas, M., & Ricci, F. (2012). Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2–3), 89–119.
- Kim, H.L., Scerri, S., Breslin, J., Decker, S., & Kim, H.G. (2008). The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In *International conference on dublin core and metadata applications*. Berlin, Germany (pp. 128–137).
- Law, E., Ahn, L., Dannenberg, R., & Crawford, M. (2007). TagATune: A game for music and sound annotation. In *Proceedings of the 8th international society for music information retrieval conference*. Vienna, Austria (pp. 361–364).
- Law, E., Settles, B., & Mitchell, T. (2010). Learning to tag from open vocabulary labels. In *European conference on machine learning and knowledge discovery in databases*. Barcelona, Spain (pp. 211–226).
- Lebret, R., Legrand, J., & Collobert, R. (2013). Is deep learning really necessary for word embeddings? In *Advances in neural information processing systems* 26. Deep learning workshop. Lake Tahoe, NV, USA (pp. 1–9).
- Levy, M., & Sandler, M. (2008). Learning latent semantic models for music from social tags. *Journal of New Music Research*, 37(2), 137–150.
- Lew, M., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1), 1–19.
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203–208.
- Mandel, M., Pascanu, R., Eck, D., Bengio, Y., Aiello, L., Schifanella, R., & Menczer, F. (2011). Contextual tag inference. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 75(1), 1–18.
- Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, England: Cambridge University Press.
- Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World Wide Web*. WWW'09. Madrid, Spain (pp. 641–650).
- Marques, O., Barenholtz, E., & Charvillat, V. (2011). Context modeling in computer vision: Techniques, implications, and applications. *Multimedia Tools and Applications*, 51, 303–339.
- Marques, G., Domingues, M., Langlois, T., & Gouyon, F. (2011). Three current issues in music autotagging. In *Proceedings of the 12th international society for music information retrieval conference*. Miami, FL, USA (pp. 795–800).
- Meyer, D., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227–234.
- Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *International conference on learning representations*. Scottsdale, AZ, USA (pp. 1–12).
- Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *11th annual conference of the international speech communication association* (pp. 1045–1048). Makuhari, Chiba, Japan: ISCA.
- Miotto, R., & Lanckriet, G. (2012). A generative context model for semantic music annotation and retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, 20(4), 1096–1108.
- Müller, H., Michoux, N., Bandon, D., & Geissbühler, A. (2004). A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *International Journal of Medical Informatics*, 73, 1–23.
- Osgood, C. (1952). The nature and measurement of meaning. *Psychological Bulletin*, 49(3), 197–237.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*. Doha, Qatar (pp. 1532–1543).
- Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *IEEE 11th international conference on computer vision*. Rio de Janeiro, Brazil.
- Ranzato, M., Boureau, Y., & LeCun, Y. (2007). Sparse feature learning for deep belief networks. In *Advances in neural information processing systems* 20. Montreal, Canada (pp. 1185–1192).
- Rasiwasia, N., & Vasconcelos, N. (2012). Holistic context models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5), 902–917.
- Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2007). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1–3), 157–173.
- Salakhutdinov, R., & Hinton, G. (2007). Learning a nonlinear embedding by preserving class neighbourhood structure. In *Proceedings of the eleventh international conference on artificial intelligence and statistics*. Vol. 2, San Juan, Porto Rico (pp. 412–419).
- Schedl, M., Flexer, A., & Urbano, J. (2013). The Neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3), 523–539.
- Schmidt, M. (2005). minFunc: unconstrained differentiable multivariate optimization in Matlab. Online: <http://www.cs.ubc.ca/~schmidt/Software/minFunc.html>.
- Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24, 35–43.
- Smeulders, A., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 1349–1380.
- Sun, M., Kim, B., Kohli, P., & Savarese, S. (2013). Relating things and stuff via object property interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(7), 1–14.
- Taylor, G., Hinton, G., & Roweis, S. (2007). Modeling human motion using binary latent variables. In *Advances in neural information processing systems* 19. NIPS'06. Vol. 19. Vancouver, Canada (p. 1345).
- Teh, W., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).
- Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, 53(2), 169–191. <http://dx.doi.org/10.1023/A:1023052124951>.

- Turnbull, D., Barrington, L., & Lanckriet, G. (2008). Five approaches to collecting tags for music. In *Proceedings of the 9th international society for music information retrieval conference*. Philadelphia, PA, USA.
- Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2007). Towards musical query-by-semantic-description using the CAL500 data set. In *Proceedings of the 30th ACM SIGIR conference on research and development in information retrieval*. Amsterdam; Netherlands (pp. 439–446).
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Wang, J., Anguera, X., Chen, X., & Yang, D. (2010). Enriching music mood annotation by semantic association reasoning. In *2010 IEEE international conference on multimedia and expo*. Singapore (pp. 1445–1450).
- Xiao, J., Hays, J., Ehinger, K., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *Conference on computer vision and pattern recognition*. San Francisco, CA, USA (pp. 3485–3492).
- Yang, Y., Huang, Z., Shen, H., & Zhou, X. (2010). Mining multi-tag association for image tagging. *World Wide Web*, 14(2), 133–156.