Probabilistic Models of Cross-Lingual Semantic Similarity in Context Based on Latent Cross-Lingual Concepts Induced from Comparable Data

Ivan Vulić and Marie-Francine Moens
Department of Computer Science
KU Leuven, Belgium
`{ivan.vulic|marie-francine.moens}@cs.kuleuven.be`

Abstract

We propose the first probabilistic approach to modeling cross-lingual semantic similarity (CLSS) in context which requires only comparable data. The approach relies on an idea of projecting words and sets of words into a shared latent semantic space spanned by language-pair independent latent semantic concepts (e.g., cross-lingual topics obtained by a multilingual topic model). These latent cross-lingual concepts are induced from a comparable corpus without any additional lexical resources. Word meaning is represented as a probability distribution over the latent concepts, and a change in meaning is represented as a change in the distribution over these latent concepts. We present new models that modulate the isolated out-of-context word representations with contextual knowledge. Results on the task of suggesting word translations in context for 3 language pairs reveal the utility of the proposed contextualized models of cross-lingual semantic similarity.

1 Introduction

Cross-lingual semantic similarity (CLSS) is a metric that measures to which extent words (or more generally, text units) describe similar semantic concepts and convey similar meanings across languages. Models of cross-lingual similarity are typically used to automatically induce bilingual lexicons and have found numerous applications in information retrieval (IR), statistical machine translation (SMT) and other natural language processing (NLP) tasks. Within the IR framework, the output of the CLSS models is a key resource in the models of dictionary-based cross-lingual information retrieval (Ballesteros and Croft, 1997; Lavrenko et al., 2002; Levow et al., 2005; Wang and Oard, 2006) or may be utilized in query expansion in cross-lingual IR models (Adriani and van Rijsbergen, 1999; Vulić et al., 2013). These CLSS models may also be utilized as an additional source of knowledge in SMT systems (Och and Ney, 2003; Wu et al., 2008). Additionally, the models are a crucial component in the cross-lingual tasks involving a sort of cross-lingual knowledge transfer, where the knowledge about utterances in one language may be transferred to another. The utility of the transfer or annotation projection by means of bilingual lexicons obtained from the CLSS models has already been proven in various tasks such as semantic role labeling (Padó and Lapata, 2009; van der Plas et al., 2011), parsing (Zhao et al., 2009; Durrett et al., 2012; Täckström et al., 2013b), POS tagging (Yarowsky and Ngai, 2001; Das and Petrov, 2011; Täckström et al., 2013a; Ganchev and Das, 2013), verb classification (Merlo et al., 2002), inducing selectional preferences (Peirsman and Padó, 2010), named entity recognition (Kim et al., 2012), named entity segmentation (Ganchev and Das, 2013), etc.

The models of cross-lingual semantic similarity from parallel corpora rely on word alignment models (Brown et al., 1993; Och and Ney, 2003), but due to a relative scarceness of parallel texts for many language pairs and domains, the models of cross-lingual similarity from comparable corpora have gained much attention recently.

All these models from parallel and comparable corpora provide *ranked lists of semantically similar words* in the target language *in isolation* or *invariably*, that is, they do not explicitly iden-

tify and encode different senses of words. In practice, it means that, given the sentence *"The coach of his team was not satisfied with the game yesterday."*, these context-insensitive models of similarity are not able to detect that the Spanish word *entrenador* is more similar to the polysemous word *coach* in the context of this sentence than the Spanish word *autocar*, although *autocar* is listed as the most semantically similar word to *coach* globally/invariably without any observed context. In another example, while Spanish words *partido*, *encuentro*, *cerilla* or *correspondencia* are all highly similar to the ambiguous English word *match* when observed in isolation, given the Spanish sentence *"She was unable to find a match in her pocket to light up a cigarette."*, it is clear that the strength of semantic similarity should change in context as only *cerilla* exhibits a strong semantic similarity to *match* within this particular sentential context.

Following this intuition, in this paper we investigate models of cross-lingual semantic similarity in context. The context-sensitive models of similarity target to *re-rank* the lists of semantically similar words based on the co-occurring contexts of words. Unlike prior work (e.g., (Ng et al., 2003; Prior et al., 2011; Apidianaki, 2011)), *we explore these models in a particularly difficult and minimalist setting that builds only on co-occurrence counts and latent cross-lingual semantic concepts induced directly from comparable corpora, and which does not rely on any other resource (e.g., machine-readable dictionaries, parallel corpora, explicit ontology and category knowledge).* In that respect, the work reported in this paper extends the current research on purely statistical data-driven distributional models of cross-lingual semantic similarity that are built upon the idea of latent cross-lingual concepts (Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011; Vulić et al., 2011; Vulić and Moens, 2013) induced from non-parallel data. While all the previous models in this framework are context-insensitive models of semantic similarity, we demonstrate how to build context-aware models of semantic similarity within the same probabilistic framework which relies on the same shared set of latent concepts.

The main contributions of this paper are:

- We present a new probabilistic approach to modeling cross-lingual semantic similarity in context based on latent cross-lingual seman-

tic concepts induced from non-parallel data.
- We show how to use the models of cross-lingual semantic similarity in the task of suggesting word translations in context.
- We provide results for three language pairs which demonstrate that contextualized models of similarity significantly outscore context-insensitive models.

2 Towards Cross-Lingual Semantic Similarity in Context

Latent Cross-Lingual Concepts. Latent cross-lingual concepts/senses may be interpreted as language-independent semantic concepts present in a multilingual corpus (e.g., document-aligned Wikipedia articles in English, Spanish and Dutch) that have their language-specific representations in different languages. For instance, having a multilingual collection in English, Spanish and Dutch, and then discovering a latent semantic concept on *Soccer*, that concept would be represented by words (actually probabilities over words $P(w|z_k)$, where w denotes a word, and z_k denotes k-th latent concept): {*player, goal, coach, ...*} in English, *balón (ball), futbolista (soccer player), equipo (team), ...*} in Spanish, and {*wedstrijd (match), elftal (soccer team), doelpunt (goal), ...*} in Dutch. Given a multilingual corpus $\mathcal{C}$, the goal is to learn and extract a set $\mathcal{Z}$ of K latent cross-lingual concepts {$z_1, ..., z_K$} that optimally describe the observed data, that is, the multilingual corpus $\mathcal{C}$. Extracting cross-lingual concepts actually implies learning *per-document concept distributions* for each document in the corpus, and discovering language-specific representations of these concepts given by *per-concept word distributions* in each language.

$\mathcal{Z} = \{z_1, ..., z_K\}$ represents the set of K latent cross-lingual concepts present in the multilingual corpus. These K semantic concepts actually span a latent cross-lingual semantic space. Each word w, irrespective of its actual language, may be represented in that latent semantic space as a K-dimensional vector, where each vector component is a conditional concept score $P(z_k|w)$.

A number of models may be employed to induce the latent concepts. For instance, one could use cross-lingual Latent Semantic Indexing (Dumais et al., 1996), probabilistic Principal Component Analysis (Tipping and Bishop, 1999), or a probabilistic interpretation of non-negative matrix

factorization (Lee and Seung, 1999; Gaussier and Goutte, 2005; Ding et al., 2008) on concatenated documents in aligned document pairs. Other more recent models include matching canonical correlation analysis (Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011) and multilingual probabilistic topic models (Ni et al., 2009; De Smet and Moens, 2009; Mimno et al., 2009; Boyd-Graber and Blei, 2009; Zhang et al., 2010; Fukumasu et al., 2012).

Due to its inherent language pair independent nature and state-of-the-art performance in the tasks such as bilingual lexicon extraction (Vulić et al., 2011) and cross-lingual information retrieval (Vulić et al., 2013), the description in this paper relies on the multilingual probabilistic topic modeling (MuPTM) framework. We draw a direct parallel between latent cross-lingual concepts and latent cross-lingual topics, and we present the framework from the MuPTM perspective, but the proposed framework is generic and allows the usage of all other models that are able to compute probability scores $P(z_k|w)$. These scores in MuPTM are induced from their output language-specific *per-topic word distributions*. The multilingual probabilistic topic models output probability scores $P(w_i^S|z_k)$ and $P(w_j^T|z_k)$ for each $w_i^S \in V^S$ and $w_j^T \in V^T$ and each $z_k \in \mathcal{Z}$, and it holds $\sum_{w_i^S \in V^S} P(w_i^S|z_k) = 1$ and $\sum_{w_j^T \in V^T} P(w_j^T|z_k) = 1$. The scores are then used to compute scores $P(z_k|w_i^S)$ and $P(z_k|w_j^T)$ in order to represent words from the two different languages in the same latent semantic space in a uniform way.

Context-Insensitive Models of Similarity. Without observing any context, the standard models of semantic word similarity that rely on the semantic space spanned by latent cross-lingual concepts in both monolingual (Dinu and Lapata, 2010a; Dinu and Lapata, 2010b) and multilingual settings (Vulić et al., 2011) typically proceed in the following manner. Latent language-independent concepts (e.g., cross-lingual topics or latent word senses) are estimated on a large corpus. The K-dimensional vector representation of the word $w_1^S \in V^S$ is:

$$vec(w_1^S) = [P(z_1|w_1^S), \dots, P(z_K|w_1^S)] \quad (1)$$

Similarly, we are able to represent any target language word w_2^T in the same latent semantic space by a K-dimensional vector with scores $P(z_k|w_2^T)$.

Each word regardless of its language is represented as a distribution over K latent concepts. The similarity between w_1^S and some word $w_2^T \in V^T$ is then computed as the similarity between their K-dimensional vector representations using some of the standard similarity measures (e.g., the Kullback-Leibler or the Jensen-Shannon divergence, the cosine measure). These methods use only global co-occurrence statistics from the training set and do not take into account any contextual information. They provide only *out-of-context word representations* and are therefore able to deliver only *context-insensitive models of similarity*.

Defining Context. Given an occurrence of a word w_1^S, we build its context set $Con(w_1^S) = \{cw_1^S, \dots, cw_r^S\}$ that comprises r words from V^S that co-occur with w_1^S in a defined *contextual scope* or granularity. In this work we do not investigate the influence of the context scope (e.g., document-based, paragraph-based, window-based contexts). Following the recent work from Huang et al. (2012) in the monolingual setting, we limit the contextual scope to the *sentential context*. However, we emphasize that the proposed models are designed to be fully functional regardless of the actual chosen context granularity. e.g., when operating in the sentential context, $Con(w_1^S)$ consists of words occurring in the same sentence with the particular instance of w_1^S. Following Mitchell and Lapata (2008), for the sake of simplicity, we impose the *bag-of-words* assumption, and do not take into account the order of words in the context set as well as context words' dependency relations to w_1^S. Investigating different context types (e.g., dependency-based) is a subject of future work.

By using all words occurring with w_1^S in a context set (e.g., a sentence) to build the set $Con(w_1^S)$, we do not make any distinction between "informative and "uninformative" context words. However, some context words bear more contextual information about the observed word w_1^S and are stronger indicators of the correct word meaning in that particular context. For instance, in the sentence *"The coach of his team was not satisfied with the game yesterday"*, words *game* and *team* are strong clues that *coach* should be translated as *entrenador* while the context word *yesterday* does not bring any extra contextual information that could resolve the ambiguity.

Therefore, in the final context set $Con(w_1^S)$ it is useful to retain only the context words that re-

ally bring extra semantic information. We achieve that by exploiting the same latent semantic space to provide the similarity score between the observed word w_1^S and each word $cw_i^S, i = 1, \ldots, r$ from its context set $Con(w_1^S)$. Each word cw_i^S may be represented by its vector $vec(cw_i^S)$ (see eq. (1)) in the same latent semantic space, and there we can compute the similarity between its vector and $vec(w_1^S)$. We can then sort the similarity scores for each cw_i^S and retain only the top scoring M context words in the final set $Con(w_1^S)$. The procedure of *context sorting and pruning* should improve the semantic cohesion between w_1^S and its context since only informative context features are now present in $Con(w_1^S)$, and we reduce the noise coming from uninformative contextual features that are not semantically related to w_1^S. Other options for the context sorting and pruning are possible, but the main goal in this paper is to illustrate the core utility of the procedure.

3 Cross-Lingual Semantic Similarity in Context via Latent Concepts

Representing Context. The probabilistic framework that is supported by latent cross-lingual concepts allows for having the K-dimensional vector representations in the same latent semantic space spanned by cross-lingual topics for: (1) Single words regardless of their actual language, and (2) Sets that comprise multiple words. Therefore, *we are able to project the observed source word, all target words, and the context set of the observed source word to the same latent semantic space spanned by latent cross-lingual concepts.*

Eq. (1) shows how to represent single words in the latent semantic space. Now, we present a way to address compositionality, that is, we show how to build the same representations in the same latent semantic space beyond the word level. We need to compute a conditional concept distribution for the context set $Con(w_1^S)$, that is, we have to compute the probability scores $P(z_k|Con(w_1^S))$ for each $z_k \in \mathcal{Z}$. Remember that the context $Con(w_1^S)$ is actually a set of r (or M after pruning) words $Con(w_1^S) = \{cw_1^S, \ldots, cw_r^S\}$. Under the *single-topic assumption* (Griffiths et al., 2007) and following Bayes' rule, it holds:

$$P(z_k|Con(w_1^S)) = \frac{P(Con(w_1^S)|z_k)P(z_k)}{P(Con(w_1^S))}$$

$$= \frac{P(cw_1^S, \ldots, cw_r^S|z_k)P(z_k)}{\sum_{l=1}^{K} P(cw_1^S, \ldots, cw_r^S|z_l)P(z_l)} \quad (2)$$

$$= \frac{\prod_{j=1}^{r} P(cw_j^S|z_k)P(z_k)}{\sum_{l=1}^{K} \prod_{j=1}^{r} P(cw_j^S|z_l)P(z_l)} \quad (3)$$

Note that here we use a simplification where we assume that all $cw_j^S \in Con(w_1^S)$ are conditionally independent given z_k. The assumption of the conditional independence of unigrams is a standard heuristic applied in *bag-of-words* model in NLP and IR (e.g., one may observe a direct analogy to probabilistic language models for IR where the assumption of independence of query words is imposed (Ponte and Croft, 1998; Hiemstra, 1998; Lavrenko and Croft, 2001)), but we have to forewarn the reader that in general the equation $P(cw_1^S, \ldots, cw_r^S|z_k) = \prod_{j=1}^{r} P(cw_j^S|z_k)$ is not exact. However, by adopting the conditional independence assumption, in case of the uniform topic prior $P(z_k)$ (i.e., we assume that we do not posses any prior knowledge about the importance of latent cross-lingual concepts in a multilingual corpus), eq. (3) may be further simplified:

$$P(z_k|Con(w_1^S)) \approx \frac{\prod_{j=1}^{r} P(cw_j^S|z_k)}{\sum_{l=1}^{K} \prod_{j=1}^{r} P(cw_j^S|z_l)} \quad (4)$$

The representation of the context set in the latent semantic space is then:

$$vec(Con(w_1^S)) = [P(z_1|Con(w_1^S)), \ldots, P(z_K|Con(w_1^S))]$$

We can then compute the similarity between words and sets of words given in the same latent semantic space in a uniform way, irrespective of their actual language. We use all these properties when building our context-sensitive CLSS models.

One remark: As a by-product of our modeling approach, by this procedure for computing representations for sets of words, we have in fact paved the way towards compositional cross-lingual models of similarity which rely on latent cross-lingual concepts. Similar to compositional models in monolingual settings (Mitchell and Lapata, 2010; Rudolph and Giesbrecht, 2010; Baroni and Zamparelli, 2010; Socher et al., 2011; Grefenstette and Sadrzadeh, 2011; Blacoe and Lapata, 2012; Clarke, 2012; Socher et al., 2012) and multilingual settings (Hermann and Blunsom, 2014; Kočiský et al., 2014), the representation of a set of words (e.g., a phrase or a sentence) is exactly the same as the representation of a single word; it is simply a K-dimensional real-valued vector. Our work on inducing structured representations of words and

text units beyond words is similar to (Klementiev et al., 2012; Hermann and Blunsom, 2014; Kočiský et al., 2014), but unlike them, we do not need high-quality sentence-aligned parallel data to induce bilingual text representations. Moreover, this work on compositionality in multilingual settings is only preliminary (e.g., we treat phrases and sentences as bags-of-words), and in future work we will aim to include syntactic information in the composition models as already done in monolingual settings (Socher et al., 2012; Hermann and Blunsom, 2013).

Intuition behind the Approach. Going back to our novel CLSS models in context, these models rely on the representations of words and their contexts in the same latent semantic space spanned by latent cross-lingual concepts/topics. The models differ in the way the contextual knowledge is fused with the out-of-context word representations.

The key idea behind these models is to represent a word w_1^S in the latent semantic space as a distribution over the latent cross-lingual concepts, but now with an additional modulation of the representation after taking its local context into account. The modulated word representation in the semantic space spanned by K latent cross-lingual concepts is then:

$$vec(w_1^S, Con(w_1^S)) = [P'(z_1|w_1^S), \ldots, P'(z_K|w_1^S)] \quad (5)$$

where $P'(z_K|w_1^S)$ denotes the recalculated (or modulated) probability score for the conditional concept/topic distribution of w_1^S after observing its context $Con(w_1^S)$. For an illustration of the key idea, see fig. 1. The intuition is that the context helps to disambiguate the true meaning of the occurrence of the word w_1^S. In other words, after observing the context of the word w_1^S, fewer latent cross-lingual concepts will share most of the probability mass in the modulated context-aware word representation.

Model I: Direct-Fusion. The first approach makes the conditional distribution over latent semantic concepts directly dependent on both word w_1^S and its context $Con(w_1^S)$. The probability score $P'(z_k|w_1^S)$ from eq. (5) for each $z_k \in \mathcal{Z}$ is then given as $P'(z_k|w_1^S) = P(z_k|w_1^S, Con(w_1^S))$.

We have to estimate the probability $P(z_k|w_1^S, Con(w_1^S))$, that is, the probability that word w_1^S is assigned to the latent concept/topic z_k given its context $Con(w_1^S)$:

$$P(z_k|w_1^S, Con(w_1^S)) = \frac{P(z_k, w_1^S)P(Con(w_1^S)|z_k)}{\sum_{l=1}^{K} P(z_l, w_1^S)P(Con(w_1^S)|z_l)} \quad (6)$$

Since $P(z_k, w_1^S) = P(w_1^S|z_k)P(z_k)$, if we closely follow the derivation from eq. (3) which shows how to project context into the latent semantic space (and again assume the uniform topic prior $P(z_k)$), we finally obtain the following formula:

$$P'(z_k|w_1^S) \approx \frac{P(w_1^S|z_k)\prod_{j=1}^{r} P(cw_j^S|z_k)}{\sum_{l=1}^{K} P(w_1^S|z_l)\prod_{j=1}^{r} P(cw_j^S|z_l)} \quad (7)$$

The ranking of all words $w_2^T \in V^T$ according to their similarity to w_1^S may be computed by detecting the similarity score between their representation in the K-dimensional latent semantic space and the modulated source word representation as given by eq. (5) and eq. (7) using any of the existing similarity functions (Lee, 1999; Cha, 2007). The similarity score $Sim(w_1^S, w_2^T, Con(w_1^S))$ between some $w_2^T \in V^T$ represented by its vector $vec(w_2^T)$ and the observed word w_1^S given its context $Con(w_1^S)$ is computed as:

$$sim(w_1^S, w_2^T, Con(w_1^S))$$
$$= SF\Big(vec(w_1^S, Con(w_1^S)), vec(w_2^T)\Big) \quad (8)$$

where SF denotes a similarity function. Words are then ranked according to their respective similarity scores and the best scoring candidate may be selected as the best translation of an occurrence of the word w_1^S given its local context. Since the contextual knowledge is integrated directly into the estimation of probability $P(z_k|w_1^S, Con(w_1^S))$, we name this context-aware CLSS model the *Direct-Fusion* model.

Model II: Smoothed-Fusion. The next model follows the modeling paradigm established within the framework of language modeling (LM), where the idea is to "back off" to a lower order N-gram in case we do not possess any evidence about a higher-order N-gram (Jurafsky and Martin, 2000). The idea now is to smooth the representation of a word in the latent semantic space induced only by the words in its local context with the out-of-context type-based representation of that word induced directly from a large training corpus. In other words, the modulated probability score $P'(z_k|w_1^S)$ from eq. (5) is calculated as:

$$P'(z_k|w_1^S) = \lambda_1 P(z_k|Con(w_1^S)) + (1 - \lambda_1)P(z_k|w_1^S) \quad (9)$$

where λ_1 is the interpolation parameter, $P(z_k|w_1^S)$ is the out-of-context conditional concept probability score as in eq. (1), and $P(z_k|Con(w_1^S))$ is given by eq. (3). This model compromises between the pure contextual word representation and
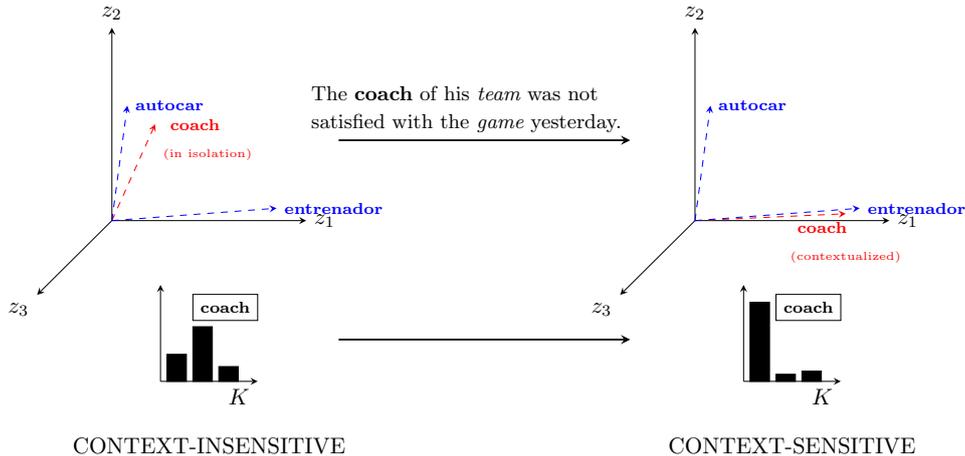
Figure 1: An illustrative toy example of the main intuitions in our probabilistic framework for building context sensitive models with only three latent cross-lingual concepts (axes z_1, z_2 and z_3): A change in meaning is reflected as a change in a probability distribution over latent cross-lingual concepts that span a shared latent semantic space. A change in the probability distribution may then actually steer an English word *coach* towards its correct (Spanish) meaning in context.

the out-of-context word representation. In cases when the local context of word w_1^S is informative enough, the factor $P(z_k|Con(w_1^S))$ is sufficient to provide the ranking of terms in V^T, that is, to detect words that are semantically similar to w_1^S based on its context. However, if the context is not reliable, we have to smooth the pure context-based representation with the out-of-context word representation (the factor $P(z_k|w_1^S)$). We call this model the *Smoothed-Fusion* model.

The ranking of words $w_2^T \in V^T$ then finally proceeds in the same manner as in Direct-Fusion following eq. (8), but now using eq. (9) for the modulated probability scores $P'(z_k|w_1^S)$.

Model III: Late-Fusion. The last model is conceptually similar to Smoothed-Fusion, but it performs smoothing at a later stage. It proceeds in two steps: (1) Given a target word $w_2^T \in V^T$, the model computes similarity scores separately between (i) the context set $Con(w_1^S)$ and w_2^T, and (ii) the word w_1^S in isolation and w_2^T (again, on the type level); (2) It linearly combines the obtained similarity scores. More formally, we may write:

$$
\begin{aligned}
&Sim(w_1^S, w_2^T, Con(w_1^S)) \\
&= \lambda_2 SF\Big(vec(Con(w_1^S)), vec(w_2^T)\Big) \\
&\quad + (1-\lambda_2)SF\Big(vec(w_1^S), vec(w_2^T)\Big)
\end{aligned}
\tag{10}
$$

where λ_2 is the interpolation parameter. Since this model computes the similarity with each target word separately for the source word in isolation and its local context, and combines the ob-

tained similarity scores after the computations, this model is called *Late-Fusion*.

4 Experimental Setup

Evaluation Task: Suggesting Word Translations in Context. Given an occurrence of a polysemous word $w_1^S \in V^S$ in the source language L_S with vocabulary V^S, the task is to choose the correct translation in the target language L_T of that particular occurrence of w_1^S from the given set $\mathcal{T} = \{t_1^T, \ldots, t_q^T\}$, $\mathcal{T} \subseteq V^T$, of its q possible translations/meanings (i.e., its translation or sense inventory). The task of *suggesting a word translation in context* may be interpreted as ranking the q translations with respect to the observed local context $Con(w_1^S)$ of the occurrence of the word w_1^S. The best scoring translation candidate in the ranked list is then the suggested correct translation for that particular occurrence of w_1^S after observing its local context $Con(w_1^S)$.

Training Data. We use the following corpora for inducing latent cross-lingual concepts/topics, i.e., for training our multilingual topic model: (i) a collection of $13,696$ Spanish-English Wikipedia article pairs (Wiki-ES-EN), (ii) a collection of $18,898$ Italian-English Wikipedia article pairs, (iii) a collection of $7,612$ Dutch-English Wikipedia article pairs (Wiki-NL-EN), and (iv) the Wiki-NL-EN corpus augmented with $6,206$ Dutch-English document pairs from Europarl (Koehn, 2005) (Wiki+EP-NL-EN). The corpora were previously used in (Vulić and Moens, 2013). No explicit use is made of sentence-level alignments in Europarl.

Sentence in Italian	Correct Translation (EN)
1. I primi **calci** furono prodotti in legno ma recentemente...	stock
2. In caso di osteoporosi si verifica un eccesso di rilascio di **calcio** dallo scheletro...	calcium
3. La crescita del **calcio** femminile professionistico ha visto il lancio di competizioni...	football
4. Il **calcio** di questa pistola (Beretta Modello 21a, calibro .25) ha le guancette in materiale...	stock

Table 1: Example sentences from our IT evaluation dataset with corresponding correct translations.

Spanish	Italian	Dutch
Ambiguous word (Possible senses/translations)	*Ambiguous word* (Possible senses/translations)	*Ambiguous word* (Possible senses/translations)
1. estación (station; season)	*1. raggio* (ray; radius; spoke)	*1. toren* (rook; tower)
2. ensayo (essay; rehearsal; trial)	*2. accordo* (chord; agreement)	*2. beeld* (image; statue)
3. núcleo (core; kernel; nucleus)	*3. moto* (motion; motorcycle)	3. *blade* (blade; leaf; magazine)
4. vela (sail; candle)	*4. calcio* (calcium; football; stock)	*4.fusie* (fusion; merger)
5. escudo (escudo; escutcheon; shield)	*5. terra* (earth; land)	*5. stam* (stem; trunk; tribe)
6. papa (Pope; potato)	*6. tavola* (board; panel; table)	*6. koper* (copper; buyer)
7. cola (glue; coke; tail; queue)	*7. campione* (champion; sample)	*7. bloem* (flower; flour)
8. cometa (comet; kite)	*8. carta* (card; paper; map)	*8. spanning* (voltage; tension; stress)
9. disco (disco; discus; disk)	*9. piano* (floor; plane; plan; piano)	*9. noot* (note; nut)
10. banda (band; gang; strip)	*10. disco* (disco; discus; disk)	*10. akkoord* (chord; agreement)
11. cinta (ribbon; tape)	*11. istruzione* (education; instruction)	*11. munt* (coin; currency; mint)
12. banco (bank; bench; shoal)	*12. gabinetto* (cabinet; office; toilet)	*12. pool* (pole; pool)
13. frente (forehead; front)	*13. torre* (rook; tower)	*13. band* (band; tyre; tape)
14. fuga (escape; fugue; leak)	*14. campo* (camp; field)	*14. kern* (core; kernel; nucleus)
15. gota (gout; drop)	*15. gomma* (rubber; gum; tyre)	*15. kop* (cup; head)

Table 2: Sets of 15 ambiguous words in Spanish, Italian and Dutch from our test set accompanied by the sets of their respective possible senses/translations in English.

All corpora are theme-aligned comparable corpora, i.e, the aligned document pairs discuss similar themes, but are in general not direct translations (except for Europarl). By training on Wiki+EP-NL-EN we want to test how the training corpus of higher quality affects the estimation of latent cross-lingual concepts that span the shared latent semantic space and, consequently, the overall results in the task of suggesting word translations in context. Following prior work (Koehn and Knight, 2002; Haghighi et al., 2008; Prochasson and Fung, 2011; Vulić and Moens, 2013), we retain only nouns that occur at least 5 times in the corpus. We record lemmatized word forms when available, and original forms otherwise. We use TreeTagger (Schmid, 1994) for POS tagging and lemmatization.

Test Data. We have constructed test datasets in Spanish (ES), Italian (IT) and Dutch (NL), where the aim is to find their correct translation in English (EN) given the sentential context. We have selected 15 polysemous nouns (see tab. 2 for the list of nouns along with their possible translations) in each of the 3 languages, and have manually extracted 24 sentences (not present in the training data) for each noun that capture different meanings of the noun from Wikipedia. In order to construct datasets that are balanced across different possible translations of a noun, in case of q different translation candidates in $\mathcal{T}$ for some word w_1^S, the dataset contains exactly $24/q$ sentences for each translation from $\mathcal{T}$. In total, we have designed 360 sentences for each language

pair (ES/IT/NL-EN), 1080 sentences in total.[1]. We have used 5 extra nouns with 20 sentences each as a development set to tune the parameters of our models. As a by-product, we have built an initial repository of ES/IT/NL ambiguous words. Tab. 1 presents a small sample from the IT evaluation dataset, and illustrates the task of suggesting word translations in context.

Evaluation Procedure. Our task is to present the system a list of possible translations and let the system decide a *single most likely translation* given the word and its sentential context. Ground truth thus contains one word, that is, one correct translation for each sentence from the evaluation dataset. We have manually annotated the correct translation for the ground truth[1] by inspecting the discourse in Wikipedia articles and the interlingual Wikipedia links. We measure the performance of all models as *Top 1* accuracy (Acc_1) (Gaussier et al., 2004; Tamura et al., 2012). It denotes the number of word instances from the evaluation dataset whose top proposed candidate in the ranked list of translation candidates from $\mathcal{T}$ is exactly the correct translation for that word instance as given by ground truth over the total number of test word instances (360 in each test dataset).

Parameters. We have tuned λ_1 and λ_2 on the development sets. We set $\lambda_1 = \lambda_2 = 0.9$ for all language pairs. We use sorted context sets (see sect. 2) and perform a cut-off at $M = 3$ most descriptive context words in the sorted context sets for all models. In the following section we discuss the utility of this context sorting and pruning, as well as its influence on the overall results.

Inducing Latent Cross-Lingual Concepts. Our context-aware models are generic and allow experimentations with different models that induce latent cross-lingual semantic concepts. However, in this particular work we present results obtained by a multilingual probabilistic topic model called bilingual LDA (Mimno et al., 2009; Ni et al., 2009; De Smet and Moens, 2009). The BiLDA model is a straightforward multilingual extension of the standard LDA model (Blei et al., 2003). For the details regarding the modeling, generative story and training of the bilingual LDA model, we refer the interested reader to the aforementioned relevant literature.

We have used the Gibbs sampling procedure

(Geman and Geman, 1984) tailored for BiLDA in particular for training and have experimented with different number of topics K in the interval $300 - 2500$. Here, we present only the results obtained with $K = 2000$ for all language pairs which also yielded the best or near-optimal performance in (Dinu and Lapata, 2010b; Vulić et al., 2011). Other parameters of the model are set to the typical values according to Steyvers and Griffiths (2007): $\alpha = 50/K$ and $\beta = 0.01$. [2]

Models in Comparison. We test the performance of our Direct-Fusion, Smoothed-Fusion and Late-Fusion models, and compare their results with the context-insensitive CLSS models described in sect. 2 (*No-Context*). We provide results with two different similarity functions: (1) We have tested different SF-s (e.g., the Kullback-Leibler and the Jensen-Shannon divergence, the cosine measure) on the K-dimensional vector representations, and have detected that in general the best scores are obtained with the Bhattacharyya coefficient (*BC*) (Cha, 2007; Kazama et al., 2010), (2) Another similarity method we use is the so-called *Cue* method (Griffiths et al., 2007; Vulić et al., 2011), which models the probability that a target word t_i^T will be generated as an association response given some cue source word w_1^S. In short, the method computes the score $P(t_i^T|w_1^S) = P(t_i^T|z_k)P(z_k|w_1^S)$. We can use the scores $P(t_i^T|w_1^S)$ obtained by inputting out-of-context probability scores $P(z_k|w_1^S)$ or modulated probability scores $P'(z_k|w_1^S)$ to produce the ranking of translation candidates.

5 Results and Discussion

The performance of all the models in comparison is displayed in tab. 3. These results lead us to several conclusions:

(i) All proposed context-sensitive CLSS models suggesting word translations in context significantly outperform context-insensitive CLSS models, which are able to produce only word translations in isolation. The improvements in results when taking context into account are ob-

---

[2]We are well aware that different hyper-parameter settings (Asuncion et al., 2009; Lu et al., 2011), might have influence on the quality of learned latent cross-lingual concepts/topics and, consequently, the quality of latent semantic space, but that analysis is not the focus of this work. Additionally, we perform *semantic space pruning* (Reisinger and Mooney, 2010; Vulić and Moens, 2013). All computations are performed over the best scoring 100 cross-lingual topics according to their respective scores $P(z_k|w_i^S)$ similarly to (Vulić and Moens, 2013).

Direction:	ES→EN		IT→EN		NL→EN (Wiki)		NL→EN (Wiki+EP)	
Model	Acc_1 (SF=BC)	Acc_1 (SF=Cue)	Acc_1 (SF=BC)	Acc_1 (SF=Cue)	Acc_1 (SF=BC)	Acc_1 (SF=Cue)	Acc_1 (SF=BC)	Acc_1 (SF=Cue)
No-Context	.406	.406	.408	.408	.433	.433	.433	.433
Direct-Fusion	.617	.575	.714	.697	.603	.592	.606	.636
Smoothed-Fusion	.664	**.703***	.731	**.789***	**.669**	**.712***	**.692**	**.761***
Late-Fusion	**.675**	.667	**.742**	.728	.667	.644	.683	.722

Table 3: Results on the 3 evaluation datasets. Translation direction is ES/IT/NL→EN. The improvements of all contextualized models over non-contextualized models are statistically significant according to a chi-square statistical significance test (p<0.05). The asterisk (*) denotes significant improvements of Smoothed-Fusion over Late-Fusion using the same significance test.

served for all 3 language pairs. The large improvements in the results (i.e., we observe an average relative increase of 51.6% for the BC+Direct-Fusion combination, 64.3% for BC+Smoothed-Fusion, 64.9% for BC+Late-Fusion, 49.1% for Cue+Direct-Fusion, 76.7% for Cue+Smoothed-Fusion, and 64.5% for Cue+Late-Fusion) confirm that the local context of a word is essential in acquiring correct word translations for polysemous words, as isolated non-contextualized word representations are not sufficient.

(ii) The choice of a similarity function influences the results. On average, the Cue method as SF outperforms other standard similarity functions (e.g., Kullback-Leibler, Jensen-Shannon, cosine, BC) in this evaluation task. However, it is again important to state that *regardless of the actual choice of SF, context-aware models that modulate out-of-context word representations using the knowledge of local context outscore context-insensitive models that utilize non-modulated out-of-context representations* (with all other parameters equal).

(iii) The Direct-Fusion model, conceptually similar to a model of word similarity in context in monolingual settings (Dinu and Lapata, 2010a), is outperformed by the other two context-sensitive models. In Direct-Fusion, the observed word and its context are modeled in the same fashion, that is, the model does not distinguish between the word and its surrounding context when it computes the modulated probability scores $P'(z_k|w_1^S)$ (see eq. (7)). Unlike Direct-Fusion, the modeling assumptions of Smoothed-Fusion and Late-Fusion provide a clear distinction between the observed word w_1^S and its context $Con(w_1^S)$ and combine the out-of-context representation of w_1^S and its contextual knowledge into a smoothed LM-inspired probabilistic model. As the results reveal, that strategy leads to better overall scores. The best scores in general are obtained by Smoothed-Fusion, but it is also outperformed by Late-Fusion in several experimental runs where BC was used as SF. However, the difference in results between Smoothed-Fusion and Late-Fusion in these experimental runs is not statistically significant according to a chi-squared significance test ($p < 0.05$).

(iv) The results for Dutch-English are influenced by the quality of training data. The performance of our models of similarity is higher for models that rely on latent-cross lingual topics estimated from the data of higher quality (i.e., compare the results when trained on Wiki and Wiki+EP in tab. 3). The overall quality of our models of similarity is of course dependent on the quality of the latent cross-lingual topics estimated from training data, and the quality of these latent cross-lingual concepts is further dependent on the quality of multilingual training data. This finding is in line with a similar finding reported for the task of bilingual lexicon extraction (Vulić and Moens, 2013).

(v) Although Dutch is regarded as more similar to English than Italian or Spanish, we do not observe any major increase in the results on both test datasets for the English-Dutch language pair compared to English-Spanish/Italian. That phenomenon may be attributed to the difference in size and quality of our training Wikipedia datasets. Moreover, while the probabilistic framework proposed in this chapter is completely language pair agnostic as it does not make any language pair dependent modeling assumptions, we acknowledge the fact that all three language pairs comprise languages coming from the same phylum, that is, the Indo-European language family. Future extensions of our probabilistic modeling framework also include porting the framework to other more distant language pairs that do not share the same roots nor the same alphabet (e.g., English-Chinese/Hindi).

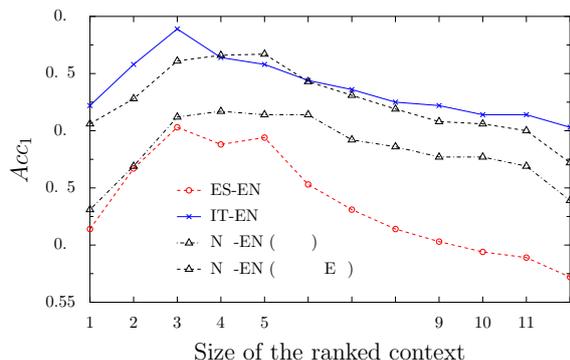**Analysis of Context Sorting and Pruning.** We

Figure 2: The influence of the size of sorted context on the accuracy of word translation in context. The model is Cue+Smoothed-Fusion.

also investigate the utility of context sorting and pruning, and its influence on the overall results in our evaluation task. Therefore, we have conducted experiments with sorted context sets that were pruned at different positions, ranging from 1 (only the most similar word to w_1^S in a sentence is included in the context set $Con(w_1^S)$) to *All* (all words occurring in a same sentence with w_1^S are included in $Con(w_1^S)$). The monolingual similarity between w_1^S and each potential context word in a sentence has been computed using BC on their out-of-context representations in the latent semantic space spanned by cross-lingual topics. Fig. 2 shows how the size of the sorted context influences the overall results. The presented results have been obtained by the Cue+Smoothed-Fusion combination, but a similar behavior is observed when employing other combinations.

Fig. 2 clearly indicates the importance of context sorting and pruning. The procedure ensures that only the most semantically similar words in a given scope (e.g., a sentence) influence the choice of a correct meaning. In other words, closely semantically similar words in the same sentence are more reliable indicators for the most probable word meaning. They are more informative in modulating the out-of-context word representations in context-sensitive similarity models. We observe large improvements in scores when we retain only the top M semantically similar words in the context set (e.g., when M=5, the scores are 0.694, 0.758, 0.717, and 0.767 for ES-EN, IT-EN, NL-EN (Wiki) and NL-EN (Wiki+EP), respectively; while the same scores are 0.572, 0.703, 0.639 and 0.672 when M=*All*).

6 Conclusions and Future Work

We have proposed a new probabilistic approach to modeling cross-lingual semantic similarity in con-

text, which relies only on co-occurrence counts and latent cross-lingual concepts which can be estimated using only comparable data. The approach is purely statistical and it does not make any additional language-pair dependent assumptions; it does not rely on a bilingual lexicon, orthographic clues or predefined ontology/category knowledge, and it does not require parallel data.

The key idea in the approach is to represent words, regardless of their actual language, as distributions over the latent concepts, and both out-of-context and contextualized word representations are then presented in the same latent space spanned by the latent semantic concepts. A change in word meaning after observing its context is reflected in a change of its distribution over the latent concepts. Results for three language pairs have clearly shown the importance of the newly developed modulated or "contextualized" word representations in the task of suggesting word translations in context.

We believe that the proposed framework is only a start, as it ignites a series of new research questions and perspectives. One may further examine the influence of context scope (e.g., document-based vs. sentence-based vs. window-based contexts), as well as context selection and aggregation (see sect. 2) on the contextualized models. For instance, similar to the model from Ó Séaghdha and Korhonen (2011) in the monolingual setting, one may try to introduce dependency-based contexts (Padó and Lapata, 2007) and incorporate the syntax-based knowledge in the context-aware CLSS modeling. It is also worth studying other models that induce latent semantic concepts from multilingual data (see sect. 2) within this framework of context-sensitive CLSS modeling. One may also investigate a similar approach to context-sensitive CLSS modeling that could operate with explicitly defined concept categories (Gabrilovich and Markovitch, 2007; Cimiano et al., 2009; Hassan and Mihalcea, 2009; Hassan and Mihalcea, 2011; McCrae et al., 2013).

References

Mirna Adriani and C. J. van Rijsbergen. 1999. Term similarity-based query expansion for cross-language information retrieval. In *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 311–322.

Marianna Apidianaki. 2011. Unsupervised cross-lingual lexical substitution. In *Proceedings of the 1st Workshop on Unsupervised Learning in NLP*, pages 13–23.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 27–34.

Lisa Ballesteros and W. Bruce Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 84–91.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1183–1193.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 546–556.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 75–82.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Sung-Hyuk Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI)*, pages 1513–1518.

Daoud Clarke. 2012. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 600–609.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 407–412.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the Web using interlingual topic modeling. In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining (SWSM@CIKM)*, pages 57–64.

Chris H. Q. Ding, Tao Li, and Wei Peng. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927.

Georgiana Dinu and Mirella Lapata. 2010a. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1162–1172.

Georgiana Dinu and Mirella Lapata. 2010b. Topic models for meaning similarity in context. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 250–258.

Susan T. Dumais, Thomas K. Landauer, and Michael Littman. 1996. Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In *Proceedings of the SIGIR Workshop on Cross-Linguistic Information Retrieval*, pages 16–23.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1–11.

Kosuke Fukumasu, Koji Eguchi, and Eric P. Xing. 2012. Symmetric correspondence topic models for multilingual text analysis. In *Proceedings of the 25th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 1295–1303.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1606–1611.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1996–2006.

Éric Gaussier and Cyril Goutte. 2005. Relation between PLSA and NMF and implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 601–602.

Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 526–533.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1394–1404.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 771–779.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1192–1201.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, pages 884–889.

Karl Moritz Hermann and Phil Blunsom. 2013. The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 894–904.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 58–68.

Djoerd Hiemstra. 1998. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 569–584.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–882.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR.

Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A Bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 247–256.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from Wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 694–702.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1459–1474.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition (ULA)*, pages 9–16.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT SUMMIT)*, pages 79–86.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 224–229.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 120–127.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 175–182.

Daniel D. Lee and H. Sebastian Seung. 1999. Algorithms for non-negative matrix factorization. In *Proceedings of the 12th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 556–562.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 25–32.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41(3):523–547.

Yue Lu, Qiaozhu Mei, and Chengxiang Zhai. 2011. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval*, 14(2):178–203.

John Philip McCrae, Philipp Cimiano, and Roman Klinger. 2013. Orthonormal explicit topic analysis for cross-lingual document matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1732–1740.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 207–214.

David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–889.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 455–462.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International World Wide Web Conference (WWW)*, pages 1155–1156.

Diarmuid Ó Séaghdha and Anna Korhonen. 2011. Probabilistic models of similarity in syntactic context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1047–1057.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 921–929.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 275–281.

Anat Prior, Shuly Wintner, Brian MacWhinney, and Alon Lavie. 2011. Translation ambiguity in and out of context. *Applied Psycholinguistics*, 32(1):93–111.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 1327–1335.

Joseph Reisinger and Raymond J. Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1182.

Sebastian Rudolph and Eugenie Giesbrecht. 2010. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 907–916.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the 24th Annual Conference on Advances in Neural*

Information Processing Systems (NIPS), pages 801–809.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1201–1211.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013a. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of ACL*, 1:1–12.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013b. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1061–1071.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 24–36.

Michael E. Tipping and Christopher M. Bishop. 1999. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 299–304.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 106–116.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 479–484.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2013. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 16(3):331–368.

Jianqiang Wang and Douglas W. Oard. 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 202–209.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 993–1000.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 200–207.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1128–1137.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 55–63.