Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses

Ivan Vulić and Marie-Francine Moens
Department of Computer Science
KU Leuven
Celestijnenlaan 200A
Leuven, Belgium
{ivan.vulic,marie-francine.moens}@cs.kuleuven.be

Abstract

We propose a new approach to identifying semantically similar words across languages. The approach is based on an idea that two words in different languages are similar if they are likely to generate similar words (which includes both source and target language words) as their top semantic word responses. Semantic word responding is a concept from cognitive science which addresses detecting most likely words that humans output as free word associations given some cue word. The method consists of two main steps: (1) it utilizes a probabilistic multilingual topic model trained on comparable data to learn and quantify the semantic word responses, (2) it provides ranked lists of similar words according to the similarity of their semantic word response vectors. We evaluate our approach in the task of bilingual lexicon extraction (BLE) for a variety of language pairs. We show that in the cross-lingual settings without any language pair dependent knowledge the response-based method of similarity is more robust and outperforms current state-of-the art methods that directly operate in the semantic space of latent cross-lingual concepts/topics.

1 Introduction

Cross-lingual semantic word similarity addresses the task of detecting words that refer to similar semantic concepts and convey similar meanings across languages. It ultimately boils down to the automatic identification of translation pairs, that is, bilingual lexicon extraction (BLE). Such lexicons and semantically similar words serve as important resources in cross-lingual knowledge induction (e.g., Zhao et al. (2009)), statistical machine translation (Och and Ney, 2003) and cross-lingual information retrieval (Ballesteros and Croft, 1997; Levow et al., 2005).

From parallel corpora, semantically similar words and bilingual lexicons are induced on the basis of word alignment models (Brown et al., 1993; Och and Ney, 2003). However, due to a relative scarceness of parallel texts for many language pairs and domains, there has been a recent growing interest in mining semantically similar words across languages on the basis of comparable data readily available on the Web (e.g., Wikipedia, news stories) (Haghighi et al., 2008; Hassan and Mihalcea, 2009; Vulić et al., 2011; Prochasson and Fung, 2011).

Approaches to detecting semantic word similarity from comparable corpora are most commonly based on an idea known as the *distributional hypothesis* (Harris, 1954), which states that words with similar meanings are likely to appear in similar contexts. Each word is typically represented by a high-dimensional vector in a feature vector space or a so-called *semantic space*, where the dimensions of the vector are its *context features*. The semantic similarity of two words, w_1^S given in the source language L_S with vocabulary V^S and w_2^T in the target language L_T with vocabulary V^T is then:

$$Sim(w_1^S, w_2^T) = SF(cv(w_1^S), cv(w_2^T)) \quad (1)$$

$cv(w_1^S) = [sc_1^S(c_1), \ldots, sc_1^S(c_N)]$ denotes a context vector for w_1^S with N context features c_k, where $sc_1^S(c_k)$ denotes the score for w_1^S associated with context feature c_k (similar for w_2^T). SF is a similarity function (e.g., cosine, the Kullback-Leibler

Proceedings of NAACL-HLT 2013, pages 106–116,
Atlanta, Georgia, 9–14 June 2013.

divergence, the Jaccard index) operating on the context vectors (Lee, 1999; Cha, 2007).

In order to compute cross-lingual semantic word similarity, one needs to design the context features of words given in two different languages that span a shared cross-lingual semantic space. Such cross-lingual semantic spaces are typically spanned by: (1) bilingual lexicon entries (Rapp, 1999; Gaussier et al., 2004; Laroche and Langlais, 2010; Tamura et al., 2012), or (2) latent language-independent semantic concepts/axes (e.g., latent cross-lingual topics) induced by an algebraic model (Dumais et al., 1996), or more recently by a generative probabilistic model (Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011; Vulić et al., 2011). Context vectors $cv(w_1^S)$ and $cv(w_2^T)$ for both source and target words are then compared in the semantic space independently of their respective languages.

In this work, we propose a new approach to constructing the shared cross-lingual semantic space that relies on a paradigm of *semantic word responding* or *free word association*. We borrow that concept from the psychology/cognitive science literature. Semantic word responding addresses a task that requires participants to produce first words that come to their mind that are related to a presented cue word (Nelson et al., 2000; Steyvers et al., 2004).

The new cross-lingual semantic space is spanned *by all vocabulary words in the source and the target language*. Each axis in the space denotes a semantic word response. The similarity between two words is then computed as the similarity between the vectors comprising their semantic word responses using any of existing SF-s. *Two words are considered semantically similar if they are likely to generate similar semantic word responses and assign similar importance to them*.

We utilize a shared semantic space of latent cross-lingual topics learned by a multilingual probabilistic topic model to obtain semantic word responses and quantify the strength of association between any cue word and its responses monolingually and across languages, and, consequently, to build *semantic response vectors*. That effectively translates the task of word similarity from the semantic space spanned by latent cross-lingual topics to the semantic space spanned by all vocabulary words in both languages.

The main contributions of this article are:

- We propose a new approach to modeling cross-lingual semantic similarity of words based on the similarity of their semantic word responses.
- We present how to estimate and quantify semantic word responses by means of a multilingual probabilistic topic model.
- We demonstrate how to employ our novel paradigm that relies on semantic word responding in the task of bilingual lexicon extraction (BLE) from comparable data.
- We show that the response-based model of similarity is more robust and obtains better results for BLE than the models that operate in the semantic space spanned by latent semantic concepts, i.e., cross-lingual topics directly.

The following sections first review relevant prior work and provide a very short introduction to multilingual probabilistic topic modeling, then describe our response-based approach to modeling cross-lingual semantic word similarity, and finally present our evaluation and results on the BLE task for a variety of language pairs.

2 Related Work

When dealing with the cross-lingual semantic word similarity, the focus of the researchers is typically on BLE, since usually the most similar words across languages are direct translations of each other. Numerous approaches emerged over the years that try to induce bilingual word lexicons on the basis of distributional information. Especially challenging is the task of mining semantically similar words from comparable data without any external knowledge source such as machine-readable seed bilingual lexicons used in (Fung and Yee, 1998; Rapp, 1999; Fung and Cheung, 2004; Gaussier et al., 2004; Morin et al., 2007; Andrade et al., 2010; Tamura et al., 2012), predefined explicit ontology or category knowledge used in (Déjean et al., 2002; Hassan and Mihalcea, 2009; Agirre et al., 2009), or orthographic clues as used in (Koehn and Knight, 2002; Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011). This work addresses that particularly difficult setting which does not assume any language pair dependent background knowledge. It makes methods

developed in such a setting applicable even on distant language pairs with scarce resources.

Recently, Griffiths et al. (2007), and Steyvers and Griffiths (2007) proposed models of free word association and semantic word similarity in the monolingual settings based on per-topic word distributions from probabilistic topic models such as pLSA (Hofmann, 1999) and LDA (Blei et al., 2003). Additionally, Vulić et al. (2011) constructed several models that utilize a shared cross-lingual topical space obtained by a multilingual topic model (Mimno et al., 2009; De Smet and Moens, 2009; Boyd-Graber and Blei, 2009; Ni et al., 2009; Jagarlamudi and Daumé III, 2010; Zhang et al., 2010) to identify potential translation candidates in the cross-lingual settings without any background knowledge. In this paper, we show that a transition from their semantic space spanned by cross-lingual topics to a semantic space spanned by all vocabulary words yields more robust models of cross-lingual semantic word similarity.

3 Modeling Word Similarity as the Similarity of Semantic Word Responses

This section contains a detailed description of our semantic word similarity method that relies on semantic word responses. Since the method utilizes the concept of multilingual probabilistic topic modeling, we first provide a very short overview of that concept, then present the intuition behind the approach, and finally describe our method in detail.

3.1 Multilingual Probabilistic Topic Modeling

Assume that we are given a *multilingual corpus* $\mathcal{C}$ of l languages, and $\mathcal{C}$ is a set of text collections $\{\mathcal{C}_1, \ldots, \mathcal{C}_l\}$ in those languages. A *multilingual probabilistic topic model* (Mimno et al., 2009; De Smet and Moens, 2009; Boyd-Graber and Blei, 2009; Ni et al., 2009; Jagarlamudi and Daumé III, 2010; Zhang et al., 2010) of a multilingual corpus $\mathcal{C}$ is defined as a set of semantically coherent multinomial distributions of words with values $P_j(w_i^j|z_k)$, $j = 1, \ldots, l$, for each vocabulary $V^1, \ldots, V^j, \ldots, V^l$ associated with text collections $\mathcal{C}_1, \ldots, \mathcal{C}_j, \ldots, \mathcal{C}_l \in \mathcal{C}$ given in languages $L_1, \ldots, L_j, \ldots, L_l$. $P_j(w_i^j|z_k)$ is calculated for each $w_i^j \in V^j$. The probability scores $P_j(w_i^j|z_k)$ build *per-topic word distributions*, and they constitute a language-specific representation (e.g., a probability value is assigned only for words from V^j) of a language-independent cross-lingual latent concept, that is, latent cross-lingual topic $z_k \in \mathcal{Z}$. $\mathcal{Z} = \{z_1, \ldots, z_K\}$ represents the set of all K latent cross-lingual topics present in the multilingual corpus. Each document in the multilingual corpus is thus considered a mixture of K cross-lingual topics from the set $\mathcal{Z}$. That mixture for some document $d_i^j \in \mathcal{C}_j$ is modeled by the probability scores $P_j(z_k|d_i^j)$ that altogether build *per-document topic distributions*.

Each cross-lingual topic from the set $\mathcal{Z}$ can be observed as a latent language-independent concept present in the multilingual corpus, but each language in the corpus uses only words from its own vocabulary to describe the content of that concept. For instance, having a multilingual collection in English, Spanish and Dutch and discovering a topic on *Soccer*, that cross-lingual topic would be represented by words (actually probabilities over words) {*player, goal, coach, ...*} in English, {*balón (ball), futbolista (soccer player), goleador (scorer), ...*} in Spanish, and {*wedstrijd (match), elftal (soccer team), doelpunt (goal), ...*} in Dutch. We have $\sum_{w_i^j \in V^j} P_j(w_i^j|z_k) = 1$, for each vocabulary V^j representing language L_j, and for each topic $z_k \in \mathcal{Z}$. Therefore, the latent cross-lingual topics also span a shared cross-lingual semantic space.

3.2 The Intuition Behind the Approach

Imagine the following thought experiment. A group of human subjects who have been raised bilingually and thus are native speakers of two languages L_S and L_T, is playing a game of word associations. The game consists of possibly an infinite number of iterations, and each iteration consists of 4 rounds. In the first round (the *S-S round*), given a word in the language L_S, the subject has to generate a list of words in the same language L_S that first occur to her/him as semantic word responses to the given word. The list is in descending order, with more prominent word responses occurring higher in the list. In the second round (the *S-T round*), the subject repeats the procedure, and generates the list of word responses to the same word from L_S, but now in the other language L_T. The third (the *T-T round*)

and the fourth round (the *T-S round*) are similar to the first and the second round, but now a list of word responses in both L_S and L_T has to be generated for some cue word from L_T. The process of generating the lists of semantic responses then continues with other cue words and other human subjects.

As the final result, for each word in the source language L_S, and each word in the target language L_T, we obtain a single list of semantic word responses comprising words in both languages. All lists are sorted in descending order, based on some association score that takes into account both the number of times a word has occurred as an associative response, as well as the position in the list in each round. We can now measure the similarity of any two words, regardless of their corresponding languages, according to the similarity of their corresponding lists that contain their word responses. Words that are equally likely to trigger the same associative responses in the human brain, and moreover assign equal importance to those responses, as provided in the lists of associative responses, are very likely to be closely semantically similar. Additionally, for a given word w_1^S in the source language L_S, some word w_2^T in L_T that has the highest similarity score among all words in L_T should be a direct word-to-word translation of w_1^S.

3.3 Modeling Semantic Word Responses via Cross-Lingual Topics

Cross-lingual topics provide a sound framework to construct a probabilistic model of the aforementioned experiment. To model semantic word responses via the shared space of cross-lingual topics, we have to set a probabilistic mass that quantifies the degree of association. Given two words $w_1, w_2 \in V^S \cup V^T$, a natural way of expressing the *asymmetric semantic association* is by modeling the probability $P(w_2|w_1)$ (Griffiths et al., 2007), that is, the probability to generate word w_2 as a response given word w_1. After the training of a multilingual topic model on a multilingual corpus, we obtain per-topic word distributions with scores $P_S(w_i^S|z_k)$ and $P_T(w_i^T|z_k)$ (see Sect. 3.1).[1] The probability $P(w_2|w_1)$ is then decomposed as follows:

$$Resp(w_1, w_2) = P(w_2|w_1) = \sum_{k=1}^{K} P(w_2|z_k)P(z_k|w_1) \quad (2)$$

The probability scores $P(w_2|z_k)$ select words that are highly descriptive for each particular topic. The probability scores $P(z_k|w_1)$ ensure that topics z_k that are semantically relevant to the given word w_1 dominate the sum, so the overall high score $Resp(w_1, w_2)$ of the semantic word response is assigned only to highly descriptive words of the semantically related topics. Using the shared space of cross-lingual topics, semantic response scores can be derived for any two words $w_1, w_2 \in V^S \cup V^T$.[1]

The generative model closely resembles the actual process in the human brain - when we generate semantic word responses, we first tend to associate that word with a related semantic/cognitive concept, in this case a cross-lingual topic (the factor $P(z_k|w_1)$), and then, after establishing the concept, we output a list of words that we consider the most prominent/descriptive for that concept (words with high scores in the factor $P(w_2|z_k)$) (Nelson et al., 2000; Steyvers et al., 2004). Due to such modeling properties, this model of semantic word responding tends to assign higher association scores for *high frequency words*. It eventually leads to *asymmetric associations/responses*. We have detected that phenomenon both monolingually and across languages. For instance, the first response to Spanish word *mutación (mutation)* is English word *gene*. Other examples include *caldera (boiler)-steam*, *deportista (sportsman)-sport*, *horario (schedule)-hour* or *pescador (fisherman)-fish*. In the other association direction, we have detected top responses such as *merchant-comercio (trade)* or *neologism-palabra (word)*. In the monolingual setting, we acquire English pairs such as *songwriter-music*, *discipline-sport*, or Spanish pairs *gripe (flu)-enfermedad (disease)*, *cuenca (basin)-río (river)*, etc.

3.4 Response-Based Model of Similarity

Eq. (2) provides a way to measure the strength of semantic word responses. In order to establish the

[1] A remark on notation throughout the paper: Since the shared space of cross-lingual topics allows us to construct a uniform representation for all words regardless of a vocabulary they belong to, due to simplicity and to stress the uniformity, we sometimes use notation $P(w_i|z_k)$ and $P(z_k|w_i)$ instead of $P_S(w_i|z_k)$ or $P_S(z_k|w_i)$ (similar for subscript T). However, the reader must be aware that, for instance, $P(w_i|z_k)$ actually means $P_S(w_i|z_k)$ if $w_i \in V^S$, and $P_T(w_i|z_k)$ if $w_i \in V^T$.

Semantic responses						Response-based similarity
dramaturgo (playwright)		play		playwright		dramaturgo
obra (play)	.101	play	.142	play	.122	playwright
escritor (writer)	.083	obra (play)	.111	escritor (writer)	.087	dramatist
play	.066	player	.033	obra (play)	.073	tragedy
writer	.050	escena (scene)	.031	writer	.060	play
poet	.047	jugador (player)	.026	poeta (poet)	.055	essayist
autor (author)	.041	adaptation	.025	poet	.053	novelist
poeta (poet)	.039	stage	.024	autor (author)	.046	drama
teatro (theatre)	.030	game	.022	teatro (theatre)	.043	tragedian
drama	.026	juego (game)	.021	tragedy	.031	satirist
contribution	.025	teatro (theatre)	.019	drama	.026	writer

Table 1: An example of top 10 semantic word responses and the final response-based similarity for some Spanish and English words. The responses are estimated from Spanish-English Wikipedia data by bilingual LDA. We can observe several interesting phenomena: (1) High-frequency words tend to appear higher in the lists of semantic responses (e.g., *play* and *obra* for all 3 words), (2) Due to the modeling properties that give preference to high-frequency words (Sect. 3.3), a word might not generate itself as the top semantic response (e.g., *playwright-play*), (3) Both source and target language words occur as the top responses in the lists, (4) Although *play* is the top semantic response in English for both *dramaturgo* and *playwright*, its list of top semantic responses is less similar to the lists of those two words, (5) Although the English word *playwright* does not appear in the top 10 semantic responses to *dramaturgo*, and *dramaturgo* does not appear in the top 10 responses to *playwright*, the more robust response-based similarity method detects that the two words are actually very similar based on their lists of responses, (6) *dramaturgo* and *playwright* have very similar lists of semantic responses which ultimately leads to detecting that *playwright* is the most semantically similar word to *dramaturgo* across the two languages (the last column), i.e., they are direct one-to-one translations of each other, (7) Another English word *dramatist* very similar to Spanish *dramaturgo* is also pushed higher in the final list, although it is not found in the list of top semantic responses to *dramaturgo*.

final similarity between two words, we have to compare their *semantic response vectors*, that is, their semantic response scores over all words in both vocabularies. The final model of word similarity closely mimics our thought experiment. First, for each word $w_i^S \in V^S$, we generate probability scores $P(w_j^S|w_i^S)$ for all words $w_j^S \in V^S$ (the *S-S* rounds). Note that $P(w_i^S|w_i^S)$ is also defined by Eq. (2). Following that, for each word $w_i^S \in V^S$, we generate probability scores $P(w_j^T|w_i^S)$, for all words $w_j^T \in V^T$ (the *S-T* rounds). Similarly, we calculate probability scores $P(w_j^T|w_i^T)$ and $P(w_j^S|w_i^T)$, for each $w_i^T, w_j^T \in V^T$, and for each $w_j^S \in V^S$ (the *T-T* and *T-S* rounds).

Now, each word $w_i \in V^S \cup V^T$ may be represented by a $(|V^S| + |V^T|)$-dimensional context vector $cv(w_i)$ as follows:[2]

$[P(w_1^S|w_i), \ldots, P(w_{|V^S|}^S|w_i), \ldots, P(w_{|V^T|}^T|w_i)]$.

We have created a language-independent cross-lingual semantic space spanned by all vocabulary words in both languages. Each feature corresponds to one word from vocabularies V^S and V^T, while the exact score for each feature in the context vector $cv(w_i)$ is precisely the probability that this word/feature will be generated as a word response given word w_i. The degree of similarity between two words is then computed on the basis of similarity between their feature vectors using some of the standard similarity functions (Cha, 2007).

The novel response-based approach of similarity removes the effect of high-frequency words that tend to appear higher in the lists of semantic word responses. Therefore, the real synonyms and translations should occur as top candidates in the lists of similar words obtained by the response-based method. That property may be exploited to identify one-to-one translations across languages and build a bilingual lexicon (see Table 1).

4 Experimental Setup

4.1 Data Collections

We work with the following corpora:

[2]We assume that the two sets V^S and V^T are disjunct. It means that, for instance, Spanish word *pie (foot)* from V^S and English word *pie* from V^T are treated as two different word types. In that case, it holds $|V^S \cup V^T| = |V^S| + |V^T|$.

- IT-EN-W: A collection of 18,898 Italian-English Wikipedia article pairs previously used by Vulić et al. (2011).
- ES-EN-W: A collection of 13,696 Spanish-English Wikipedia article pairs.
- NL-EN-W: A collection of 7,612 Dutch-English Wikipedia article pairs.
- NL-EN-W+EP: The NL-EN-W corpus augmented with 6,206 Dutch-English document pairs from Europarl (Koehn, 2005). Although Europarl is a parallel corpus, no explicit use is made of sentence-level alignments.

All corpora are theme-aligned, that is, the aligned document pairs discuss similar subjects, but are in general not direct translations (except the Europarl document pairs). NL-EN-W+EP serves to test whether better semantic responses could be learned from data of higher quality, and to measure how it affects the response-based similarity method and the quality of induced lexicons. Following (Koehn and Knight, 2002; Haghighi et al., 2008; Prochasson and Fung, 2011), we consider only noun word types. We retain only nouns that occur at least 5 times in the corpus. We record the lemmatized form when available, and the original form otherwise. Again following their setup, we use TreeTagger (Schmid, 1994) for POS tagging and lemmatization.

4.2 Multilingual Topic Model

The multilingual probabilistic topic model we use is a straightforward multilingual extension of the standard Blei et al.'s LDA model (Blei et al., 2003) called bilingual LDA (Mimno et al., 2009; Ni et al., 2009; De Smet and Moens, 2009). For the details regarding the modeling assumptions, generative story, training and inference procedure of the bilingual LDA model, we refer the interested reader to the aforementioned relevant literature. The potential of the model in the task of bilingual lexicon extraction was investigated before (Mimno et al., 2009; Vulić et al., 2011), and it was also utilized in other cross-lingual tasks (e.g., Platt et al. (2010); Ni et al. (2011)). We use Gibbs sampling for training. In a typical setting for mining semantically similar words using latent topic models in both monolingual (Griffiths et al., 2007; Dinu and Lapata, 2010) and cross-lingual setting (Vulić et al., 2011), the best results are obtained with the number of topics set to a few thousands (≈ 2000). Therefore, our bilingual LDA model on all corpora is trained with the number of topics $K = 2000$. Other parameters of the model are set to the standard values according to Steyvers and Griffiths (2007): $\alpha = 50/K$ and $\beta = 0.01$. We are aware that different hyper-parameter settings (Asuncion et al., 2009; Lu et al., 2011), might have influence on the quality of learned cross-lingual topics, but that analysis is out of the scope of this paper.

4.3 Compared Methods

We evaluate and compare the following word similarity approaches in all our experiments:
1) The method that regards the lists of semantic word responses across languages obtained by Eq. (2) directly as the lists of semantically similar words (**Direct-SWR**).
2) The state-of-the-art method that employs a similarity function (SF) on the K-dimensional word vectors $cv(w_i)$ in the semantic space of latent cross-lingual topics. The dimensions of the vectors are conditional topic distribution scores $P(z_k|w_i)$ that are obtained by the multilingual topic model directly (Steyvers and Griffiths, 2007; Vulić et al., 2011). We have tested different SF-s (e.g., the Kullback-Leibler and the Jensen-Shannon divergence, the cosine measure), and have detected that in general the best scores are obtained when using the Bhattacharyya coefficient (BC) (Bhattacharyya, 1943; Kazama et al., 2010) (**Topic-BC**).
3) The best scoring similarity method from Vulić et al. (2011) named **TI+Cue**. This state-of-the-art method also operates in the semantic space of latent cross-lingual concepts/topics.
4) The response-based similarity described in Sect. 3. As for *Topic-BC*, we again use BC as the similarity function, but now on $|V^S \cup V^T|$-dimensional context vectors in the semantic space spanned by all words in both vocabularies that represent semantic word responses (**Response-BC**). Given two N-dimensional word vectors $cv(w_1^S)$ and $cv(w_2^T)$, the BC or the *fidelity* measure (Cha, 2007) is defined as:

$$BC(cv(w_1^S), cv(w_2^T)) = \sum_{n=1}^{N} \sqrt{sc_1^S(c_n) \cdot sc_2^T(c_n)} \quad (3)$$

Corpus:	IT-EN-W			ES-EN-W			NL-EN-W			NL-EN-W+EP		
Method	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}
Direct-SWR	.501	.576	.740	.332	.437	.675	.186	.254	.423	.344	.450	.652
Topic-BC	.578	.667	.834	.433	.576	.843	**.237**	.314	.489	.534	.630	.836
TI+Cue	.597	.702	**.897**	.429	.569	.828	.225	.296	.459	.446	.569	.808
Response-BC	**.622**	**.729**	.882	**.517**	**.635**	**.891**	.236	**.320**	**.511**	**.574**	**.653**	**.864**

Table 2: BLE performance of all the methods for Italian-English, Spanish-English and Dutch-English (with 2 different corpora utilized for the training of bilingual LDA and the estimation of semantic word responses for Dutch-English).

For the *Topic-BC* method $N = K$, while $N = |V^S \cup V^T|$ for *Response-BC*. Additionally, since $P(z_k|w_i) > 0$ and $P(w_k|w_i) > 0$ for each $z_k \in \mathcal{Z}$ and each $w_k \in V^S \cup V^T$, a lot of probability mass is assigned to topics and semantic responses that are completely irrelevant to the given word. Reducing the dimensionality of the semantic representation a posteriori to only a smaller number of most important semantic axes in the semantic spaces should decrease the effects of that statistical noise, and even more firmly emphasize the latent correlation among words. The utility of such *semantic space truncating* or *feature pruning* in monolingual settings (Reisinger and Mooney, 2010) was also detected previously for LSA and LDA-based models (Landauer and Dumais, 1997; Griffiths et al., 2007). Therefore, unless noted otherwise, we perform all our calculations over the best scoring 200 cross-lingual topics and the best scoring 2000 semantic word responses.[3]

4.4 Evaluation

Ground truth translation pairs.[4] Since our task is bilingual lexicon extraction, we designed a set of ground truth one-to-one translation pairs for all 3 language pairs as follows. For Dutch-English and Spanish-English, we randomly sampled a set of Dutch (Spanish) nouns from our Wikipedia corpora. Following that, we used the *Google Translate* tool plus an additional annotator to translate those words to English. The annotator manually revised the lists and retained only words that have their corresponding translation in the English vocabulary. Additionally, only one possible translation was annotated as correct. When more than 1 translation is possible, the annotator marked as correct the translation that occurs more frequently in the English Wikipedia data. Finally, we built a set of 1000 one-to-one translation pairs for Dutch-English and Spanish-English. The same procedure was followed for Italian-English, but there we obtained the ground truth one-to-one translation pairs for 1000 most frequent Italian nouns in order to test the effect of word frequency on the quality of semantic word responses and the overall lexicon quality.

Evaluation metrics. All the methods under consideration actually retrieve ranked lists of semantically similar words that could be observed as potential translation candidates. We measure the performance on BLE as *Top M* accuracy (Acc_M). It denotes the number of source words from ground truth translation pairs whose top M semantically similar words contain the correct translation according to our ground truth over the total number of ground truth translation pairs (=*1000*) (Tamura et al., 2012). Additionally, we compute the *mean reciprocal rank (MRR)* scores (Voorhees, 1999).

5 Results and Discussion

Table 2 displays the performance of each compared method on the BLE task. It shows the difference in results for different language pairs and different corpora used to extract latent cross-lingual topics and estimate the lists of semantic word responses. Example lists of semantically similar words over all 3 language pairs are shown in Table 3. Based on these results, we are able to derive several conclusions:

(i) *Response-BC* performs consistently better than the other 3 methods over all corpora and all language pairs. It is more robust and is able to find some cross-lingual similarities omitted by the other meth-

[3]The values are set empirically. Calculating similarity $Sim(w_1^S, w_2^T)$ may be interpreted as: "Given word w_1^S detect how similar word w_2^T is to the word w_1^S." Therefore, when calculating $Sim(w_1^S, w_2^T)$, even when dealing with symmetric similarity functions such as BC, we always consider only the scores $P(\cdot|w_1^S)$ for truncating.

[4]Available online: `http://people.cs.kuleuven.be/~ivan.vulic/software/`

Italian-English (IT-EN)			Spanish-English (ES-EN)			Dutch-English (NL-EN)		
(1) **affresco** (fresco)	(2) **spigolo** (edge)	(3) **coppa** (cup)	(1) **caza** (hunting)	(2) **discurso** (speech)	(3) **comprador** (buyer)	(1) **behoud** (conservation)	(2) **schroef** (screw)	(3) **spar** (fir)
fresco	polyhedron	club	*hunting*	rhetoric	purchase	*conservation*	socket	conifer
mural	polygon	competition	hunt	oration	seller	preservation	wire	pine
nave	vertices	final	hunter	*speech*	tariff	heritage	wrap	firewood
wall	diagonal	champion	hound	discourse	market	diversity	wrench	seedling
testimonial	*edge*	football	safari	dialectic	bidding	emphasis	*screw*	weevil
apse	vertex	trophy	huntsman	rhetorician	auction	consequence	pin	chestnut
rediscovery	binomial	team	wildlife	oratory	bid	danger	fastener	acorn
draughtsman	solid	relegation	animal	wisdom	microeconomics	contribution	torque	girth
ceiling	graph	tournament	ungulate	oration	trade	decline	pipe	lumber
palace	modifier	soccer	chase	persuasion	listing	framework	routing	bark

Table 3: Example lists of top 10 semantically similar words across all 3 language pairs according to our *Response-BC* similarity method, where the correct translation word is: (col. 1) found as the most similar word, (2) contained lower in the list, and (3) not found in the top 10 words.

IT-EN	ES-EN	NL-EN
direttore-director	flauta-flute	kustlijn-coastline
radice-root	eficacia-efficacy	begrafenis-funeral
sintomo-symptom	empleo-employment	mengsel-mixture
perdita-loss	descubierta-discovery	lijm-glue
danno-damage	desalojo-eviction	kijker-viewer
battaglione-battalion	miedo-fear	oppervlak-surface

Table 4: Example translations found by the *Response-BC* method, but missed by the other 3 methods.

ods (see Table 4). The overall quality of the cross-lingual word similarities and lexicons extracted by the method is dependent on the quality of estimated semantic response vectors. The quality of these vectors is of course further dependent on the quality of multilingual training data. For instance, for Dutch-English, we may observe a rather spectacular increase in overall scores (the tests are performed over the same set of 1000 words) when we augment Wikipedia data with Europarl data (compare the scores for NL-EN-W and NL-EN-W+EP).

(ii) A transition from a semantic space spanned by cross-lingual topics (*Topic-BC*) to a semantic space spanned by vocabulary words (*Response-BC*) leads to better results over all corpora and language pairs. The difference is less visible when using training data of lesser quality (the scores for NL-EN-W). Moreover, since the shared space of cross-lingual topics is used to obtain and quantify semantic word responses, the quality of learned cross-lingual topics influences the quality of semantic word responses. If the semantic coherence of the cross-lingual topical space is unsatisfying, the method is unable to generate good semantic response vectors, and ultimately unable to correctly identify semantically similar words across languages.

(iii) Due to its modeling properties that assign more importance to high-frequency words, *Direct-SWR* produces reasonable results in the BLE task only for high-frequency words (see results for IT-EN-W). Although Eq. (2) models the concept of semantic word responding in a sound way (Griffiths et al., 2007), using the semantic word responses directly is not suitable for the actual BLE task.

(iv) The effect of word frequency is clearly visible when comparing the results obtained on IT-EN-W with the results obtained on the other Wikipedia corpora. High-frequency words produce more redundancies in training data that are captured by statistical models such as latent topic models. High-frequency words then obtain better estimates of their semantic response vectors which consequently leads to better overall scores. The effect of word frequency on statistical methods in the BLE task was investigated before (Pekar et al., 2006; Prochasson and Fung, 2011; Tamura et al., 2012), and we also confirm their findings.

(v) Unlike (Koehn and Knight, 2002; Haghighi et al., 2008), our response-based method does not rely on any orthographic features such as cognates or words shared across languages. It is a pure statistical method that only relies on word distributions over a multilingual corpus. Based on these distributions, it performs the initial shallow semantic analysis of the corpus by means of a multilingual probabilistic model. The method then builds, via the concept of semantic word responding, a language-

independent semantic space spanned by all vocabulary words/responses in both languages. That makes the method portable to distant language pairs. However, for similar languages, including more evidence such as orthographic clues might lead to further increase in scores, but we leave that for future work.

6 Conclusion

We have proposed a new statistical approach to identifying semantically similar words across languages that relies on the paradigm of semantic word responding previously defined in cognitive science. The proposed approach is robust and does not make any additional language-pair dependent assumptions (e.g., it does not rely on a seed lexicon, orthographic clues or predefined concept categories). That effectively makes it applicable to any language pair. Our experiments on the task of bilingual lexicon extraction for a variety of language pairs have proved that the response-based approach is more robust and outperforms the methods that operate in the semantic space of latent concepts (e.g., cross-lingual topics) directly.

Acknowledgments

We would like to thank Steven Bethard and the anonymous reviewers for their useful suggestions. This research has been carried out in the framework of the TermWise Knowledge Platform (IOF-KP/09/001) funded by the Industrial Research Fund, KU Leuven, Belgium.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL-HLT*, pages 19–27.

Daniel Andrade, Tetsuya Nasukawa, and Junichi Tsujii. 2010. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of COLING*, pages 19–27.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of 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 SIGIR*, pages 84–91.

A. Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:199–209.

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 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.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of ACL*, 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 *CIKM Workshop on Social Web Search and Mining (SWSM)*, pages 57–64.

Hervé Déjean, Eric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of COLING*, pages 1–7.

Georgiana Dinu and Mirella Lapata. 2010. Topic models for meaning similarity in context. In *Proceedings of 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.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of EMNLP*, pages 57–63.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING*, pages 414–420.

Eric 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 ACL*, pages 526–533.

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 ACL*, pages 771–779.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of EMNLP*, pages 1192–1201.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of SIGIR*, pages 50–57.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of ECIR*, pages 444–456.

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 ACL*, pages 247–256.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, pages 79–86.

Thomas K. Landauer and Susan T. Dumais. 1997. Solutions to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of COLING*, pages 617–625.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of 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*, 41:523–547.

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.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of EMNLP*, pages 880–889.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of ACL*, pages 664–671.

Douglas L. Nelson, Cathy L. McEvoy, and Simon Dennis. 2000. What is free association and what does it measure? *Memory and Cognition*, 28:887–899.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of WWW*, pages 1155–1156.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross lingual text classification by mining multilingual topics from Wikipedia. In *Proceedings of WSDM*, pages 375–384.

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

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

John C. Platt, Kristina Toutanova, and Wen-Tau Yih. 2010. Translingual document representations from discriminative projections. In *Proceedings of EMNLP*, pages 251–261.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Proceedings of ACL*, pages 1327–1335.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL*, pages 519–526.

Joseph Reisinger and Raymond J. Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of EMNLP*, pages 1173–1182.

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

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

Mark Steyvers, Richard M. Shiffrin, and Douglas L. Nelson. 2004. Word association spaces for predicting semantic similarity effects in episodic memory. In *Experimental Cognitive Psychology and Its Applications*, pages 237–249.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of EMNLP*, pages 24–36.

Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of TREC*, pages 77–82.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL*, pages 479–484.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of 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 ACL*, pages 55–63.