Sparse Bilingual Word Representations for Cross-lingual Lexical Entailment

Yogarshi Vyas and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
yogarshi@cs.umd.edu, marine@cs.umd.edu

Abstract

We introduce the task of *cross-lingual lexical entailment*, which aims to detect whether the meaning of a word in one language can be inferred from the meaning of a word in another language. We construct a gold standard for this task, and propose an unsupervised solution based on distributional word representations. As commonly done in the monolingual setting, we assume a word e entails a word f if the prominent context features of e are a subset of those of f. To address the challenge of comparing contexts across languages, we propose a novel method for inducing sparse bilingual word representations from monolingual and parallel texts. Our approach yields an F-score of 70%, and significantly outperforms strong baselines based on translation and on existing word representations.

1 Introduction

Multilingual Natural Language Processing lacks techniques to automatically compare and contrast the meaning of words across languages. Machine translation (Koehn, 2010) lets us discover translation correspondences in bilingual texts, but a word and its translation often do not cover the exact same semantic space: distinct word senses might translate differently (Gale et al., 1992; Diab and Resnik, 2002, among others); semantic relations and associations do not always translate, an important issue when constructing multilingual ontologies (Fellbaum and Vossen, 2012); and words in parallel text might be translated non-literally due to lexical gaps (Santos, 1990; Bentivogli and Pianta, 2000) or decisions of the translator, as becomes clear when comparing multiple translations of the same source text (Bhagat and Hovy, 2013).

As a result, correct word translations found in parallel corpora exhibit a variety of relations including equivalence, hypernymy, and meronymy. For instance, even after removing noisy examples (Johnson et al., 2007) from a Machine Translation bilexicon induced from parallel corpora (Koehn et al., 2007), we find that the French word "appartement" (apartment) is linked to related but not strictly equivalent English words, such as "home" or "condo".

house	‖	*foyer* (foyer)
house	‖	*maison* (house)
house	‖	*chambre* (chamber)
home	‖	*appartement*
condo	‖	*appartement*
apartment	‖	*appartement*

Table 1: Examples of translations drawn from an English-French bilexicon automatically learned on parallel text.

We aim to design models that capture these differences and similarities in word meaning across languages, beyond translation correspondences. As a first step, we introduce *cross-lingual lexical entailment*, the task of detecting whether the meaning of a word in one language can be inferred from the meaning of a word in another language. In monolingual settings, lexical entailment has received significant attention as a representation-agnostic way of modeling lexical semantics, and as a step toward textual inference (Zhitomirsky-Geffet and Dagan, 2009; Turney and Mohammad, 2015; Levy et

Proceedings of NAACL-HLT 2016, pages 1187–1197,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

al., 2015; Pavlick et al., 2015). We hypothesize that the cross-lingual task can help do the same with multilingual texts.

Building on prior work on the monolingual task, we take an unsupervised approach, and use a directional semantic similarity metric motivated by the distributional inclusion hypothesis (Geffet and Dagan, 2005a; Kotlerman et al., 2010): we assume a word e entails a word f if the prominent context features of e are a subset of those of f. However, we face a new challenge in the cross-lingual case: how can we represent and compare word contexts across languages? Our solution leverages recent work on sparse representations for natural language processing. We develop sparse bilingual word representations that represent contexts based on interpretable dimensions that are aligned across languages.

As we will see, this approach successfully detects cross-lingual lexical entailment (with an F-score of 70%), and significantly outperforms strong baselines based (1) on machine translation, and (2) on existing dense bilingual word representations. Along the way, we construct a new dataset to evaluate cross-lingual lexical entailment, and also show the benefits of our approach in the monolingual setting.

2 A Cross-Lingual View of Lexical Entailment

Zhitomirsky-Geffet and Dagan (2009) formalize lexical entailment as a substitutional relationship. Under their definition, given a word pair (w, v), w entails v if the following two conditions are fulfilled

1. The meaning of a possible sense of w implies a possible sense of v, and

2. w can substitute for v in a sentence, such that the meaning of the modified sentence entails the meaning of the original sentence.

As a result, monolingual lexical entailment includes various semantic relations, such as synonymy, hypernymy, some meronymy relations, but also cause-effect relations (*murder* entails *death*), and other associations (*ocean* entails *water*) (Kotlerman et al., 2010).

We extend this definition to the cross-lingual case by modifying the second condition. Given a word pair (w', v'), where w' is a word in language ***e*** and v' is a word in language ***f***, w' entails v' if

1. The meaning of a possible sense of w' implies a possible sense of v', and

2. Given a sentence T in ***f*** containing v', w' can substitute for v' in the translation of T in ***e***, such that the meaning of the modified sentence entails the meaning of the original sentence.

Cross-lingual lexical entailment helps us refine our understanding of semantic mappings across languages: while the French word *ouvrier* can be translated as *worker* in English, knowing that *worker* does not entail *ouvrier* could be useful in many multilingual applications, including machine translation and its evaluation, question answering or entity linking in multilingual texts.

As can be seen in Table 2, lexical entailment is not always preserved by translation: while *aspirin* entails the English word *drug*, it does not entail the French *drogue*, which only refers to the narcotic sense of *drug* and not to its medicinal sense.

English-English	**English-French**
affection $\rightarrow$ feeling	affection $\rightarrow$ sentiment
aspirin $\rightarrow$ drug	aspirin $\not\rightarrow$ drogue
water $\rightarrow$ wet	water $\rightarrow$ humide
feeling $\not\rightarrow$ nostalgia	feeling $\not\rightarrow$ nostalgie

Table 2: Examples of monolingual and cross-lingual lexical entailment: $\rightarrow$ can be read as "entails", $\not\rightarrow$ as "does not entail".

When evaluating lexical entailment, we use the same approach as in monolingual tasks (Baroni et al., 2012; Baroni and Lenci, 2011; Kotlerman et al., 2010; Turney and Mohammad, 2015): given a bilingual word pair, systems are asked to make a binary true/false decision on whether the first word entails the second. We describe the collection of gold standard annotations in Section 5.2.

3 Unsupervised Detection of Lexical Entailment

We choose to detect lexical entailment without supervision. As in the monolingual case, detection can be done using a scoring function which quantifies the *directional* semantic similarity of an input word pair. On monolingual tasks, despite reaching better

performance, supervised systems do not really learn entailment relations for word pairs, but instead learn when a particular word in the pair is a "prototypical hypernym" (Levy et al., 2015). [1] Thus, we limit our investigation to unsupervised models. As a result, our approach only requires a small number of annotated examples to tune the scoring threshold.

We use the *balAPinc* score (Kotlerman et al., 2009), which is based on the distributional inclusion hypothesis (Geffet and Dagan, 2005b): given feature representations of the contexts of two words u and v, u is assumed to entail v if all features of u tend to appear within the features of v.

Formally, *balAPinc* is the geometric mean of a symmetric similarity score, *LIN* (Lin, 1998), and an asymmetric score, *APinc*. Given a directional entailment pair $(u \rightarrow v)$,

$$\textit{balAPinc}(u \rightarrow v) = \sqrt{\textit{LIN}(u,v) \cdot \textit{APinc}(u \rightarrow v)}$$

Assume we are given ranked feature lists FV_u and FV_v for words u and v respectively. Let $w_u(f)$ denote the weight of a particular feature f in FV_u. *LIN* is defined by

$$\textit{LIN}(u,v) = \frac{\sum_{f \in FV_u \cap FV_v} [w_u(f) + w_v(f)]}{\sum_{f \in FV_u} w_u(f) + \sum_{f \in FV_v} w_v(f)} \quad (1)$$

APinc is a modified asymmetric version of the Average Precision metric used in Information Retrieval:

$$\textit{APinc}(u \rightarrow v) = \frac{\sum_{r=1}^{|FV_u|} [P(r, FV_u, FV_v) \cdot rel'(f_r)]}{|FV_u|} \quad (2)$$

where,

$$P(r, FV_u, FV_v) = \frac{|\text{\# features of } v \text{ in top } r \text{ features of } u\ |}{r}$$

$$rel'(f) = \begin{cases} 1 & \text{if} f \in FV_u \\ 0 & \text{otherwise} \end{cases}$$

[1] Given a word pair such as (dog,animal), supervised methods tend to learn that animal is very likely to be a category word i.e. one that is likely to be a hypernym, and do not take into account the relationship of animal with dog.

Thus, to use *balAPinc* for cross-lingual lexical entailment, we need a ranked list of features that capture information about the context of words in two languages. In the monolingual case, features are dimensions in a distributional semantic space. For the cross-lingual task, we need to represent words in two languages in the same space, or in spaces where a one-to-one mapping between dimensions exists.

4 Learning Sparse Bilingual Word Representations

As we will see in Section 9, there is a wealth of existing methods for learning representations that capture context of words in two different languages in the literature. However, they have been evaluated on tasks that do not require much semantic analysis, such as translation lexicon induction or document categorization. In contrast, detecting lexical entailment requires the ability to capture more subtle semantic distinctions. This requires bilingual representations to capture both the full range of word contexts observed in original language texts, as well as cross-lingual correspondences from translated texts.

We propose a new model that uses *sparse nonnegative embeddings* to represent word contexts as interpretable dimensions, and facilitate context comparisons across languages. This is an instance of sparse coding, which consists of modeling data vectors as sparse linear combinations of basis elements. In contrast with dimensionality reduction techniques such as PCA, the learned basis vectors need not be orthogonal, which gives more flexibility to represent the data (Mairal et al., 2009). These models have been introduced as word representations in monolingual settings (Murphy et al., 2012) with the goal of obtaining interpretable, cognitively-plausible representations. We review the monolingual models, before introducing our novel bilingual formulation.

4.1 Review: Learning Monolingual Sparse Representations

Previous work (Murphy et al., 2012; Faruqui et al., 2015) on obtaining sparse monolingual representations is based on a variant of the Nonnegative Matrix Factorization problem. Given a matrix $\mathbf{X}$ containing v dense word representations arranged row-wise, sparse representations for the v words can be ob-

tained by solving the following optimization problem

$$\begin{aligned} \underset{\mathbf{A},\mathbf{D}}{\text{argmin}} \quad & \sum_{i=1}^{v} ||\mathbf{A}_i\mathbf{D}^T - \mathbf{X}_i||_2^2 + \lambda||\mathbf{A}_i||_1 \\ \text{subject to} \quad & \mathbf{A}_{ij} >= 0, \forall i,j \\ & \mathbf{D}_i^T\mathbf{D}_i <= 1, \forall i \end{aligned} \tag{3}$$

The first term in the objective 3 aims to factorize the dense representation matrix $\mathbf{X}$ into two matrices, $\mathbf{A}$ and $\mathbf{D}$ such that the l_2 reconstruction error is minimized. The second term is an l_1 regularizer on $\mathbf{A}$ which encourages sparsity, where the level of sparsity is controlled by the λ hyperparameter. This, together with the non-negativity constraint, helps in obtaining sparsified and interpretable representations in $\mathbf{A}$ since non-negativity has been shown to correlate with interpretability. Note that the objective function on its own is degenerate since it can be trivially optimized by making the entries of $\mathbf{D}$ arbitrarily large and choosing corresponding small values as entries of $\mathbf{A}$. To avoid this, an additional constraint is imposed on $\mathbf{D}$.

4.2 Proposed Bilingual Model

Learning bilingual word representations for entailment requires two sources of information:

- Monolingual distributional representations independently learned from large amounts of text in each language. We denote them as two input matrices, $\mathbf{X_e}$ and $\mathbf{X_f}$, of respective sizes $v_e \times n_e$ and $v_f \times n_f$. Each row in $\mathbf{X_e}$ represents the representation of a particular word in the first language, ***e***, while $\mathbf{X_f}$ represents word representations for the other language ***f***.

- Cross-lingual correspondences that enable comparison across languages. We define a "score" matrix $\mathbf{S}$ of size $v_e \times v_f$, which captures high-confidence correspondences between the vocabularies of the two languages. There are many ways of defining $\mathbf{S}$. As a starting point, we define each row of $\mathbf{S}$ as a one-hot vector that identifies the word in ***f*** that is most frequently aligned with the ***e*** word for that row in a large parallel corpus. This reduction leads to a many-to-one mapping from ***e*** to ***f***, which captures translation ambiguity by allowing multiple words in ***e*** to be aligned to the same word in ***f***.

We formulate the following optimization problem to obtain sparse bilingual representations:

$$\begin{aligned} \underset{\mathbf{A_e},\mathbf{D_e},\mathbf{A_f},\mathbf{D_f}}{\text{argmin}} & \sum_{i=i}^{v_e} \frac{1}{2}||\mathbf{A_e}_i\mathbf{D_e}^{\mathrm{T}} - \mathbf{X_e}_i||_2^2 + \lambda_e||\mathbf{A_e}_i||_1 \\ & + \sum_{j=i}^{v_f} \frac{1}{2}||\mathbf{A_f}_j\mathbf{D_f}^{\mathrm{T}} - \mathbf{X_f}_j||_2^2 + \lambda_f||\mathbf{A_f}_j||_1 \\ & + \sum_{i=1}^{v_e}\sum_{j=1}^{v_f} \frac{1}{2}\lambda_x\mathbf{S}_{ij}||\mathbf{A_e}_i - \mathbf{A_f}_j||_2^2 \\ \text{subject to} \quad & \mathbf{A_e} > 0 \,; {\mathbf{D_e}_i}^{\mathrm{T}}.\mathbf{D_e}_i \leq 1, 1 \leq i \leq v_e; \\ & \mathbf{A_f} > 0 \,; {\mathbf{D_f}_j}^{\mathrm{T}}.\mathbf{D_f}_j \leq 1, 1 \leq j \leq v_f; \end{aligned} \tag{4}$$

The first two rows and the constraints in Equation 4 can be understood as in Equation 3 - they encourage sparsity in word representations for each language. The third row imposes bilingual correspondence constraints, weighted by the regularizer λ_x: it encourages words in ***e*** and ***f*** that are strongly aligned according to $\mathbf{S}$ to have similar representations.

4.3 Optimization

Equations 3 and 4 define non-differentiable, non-convex optimization problems and finding the globally optimally solution is not feasible. However, various methods used to solve convex problems work well in practice. We use *Forward Backward Splitting*, a proximal gradient method for which an efficient generic solver, FASTA, is available (Goldstein et al., 2015; Goldstein et al., 2014). FASTA (Fast Adaptive Shrinkage / Thresholding Algorithm) is designed to minimize functions of the form $f(Ax) + g(x)$, where f is a differentiable function, g is a function (possibly non-differentiable) for which we can calculate the proximal operator, and A is a linear operator. For the objective function in our model, the l_1 terms form g and the l_2 terms form f.

We have now described all components of the model required to detect bilingual lexical entailment: solving objective 4 as described yields sparse representations for words in the two languages that can be compared directly using the *balAPinc* metric.

5 Constructing a Gold Standard

5.1 Existing Monolingual Test Suites

A comprehensive suite of lexical entailment test beds is available for English (Levy et al., 2015). They were constructed either by asking humans to annotate entailment relations directly (Kotlerman et al., 2010), or by deriving entailment labels from semantic relation annotations (Baroni et al., 2012; Baroni and Lenci, 2011; Turney and Mohammad, 2015). Each test set has 900 to 1300 positive examples of lexical entailment - word pairs (w, v) such that $w \rightarrow v$. All but one are balanced.

5.2 Creating a Cross-Lingual Test Set

We select French as the second language: it is a good starting point for studying cross-lingual entailment, as it is a resource-rich language with many available bilingual annotators. We will construct data sets for more distant language pairs in future work.

We aim to construct a balanced test set of positive and negative bilingual entailment examples in the spirit of the existing English test beds. While it is attractive to leverage existing English examples, we cannot translate them directly as entailment relations might be affected by translation ambiguity (as illustrated in Table 2).

We therefore obtain annotated bilingual examples using a two step process: (1) automatic translation of monolingual examples, and (2) manual annotation through crowdsourcing. For a sample of positive examples of entailment $w_e \rightarrow v_e$ in the monolingual datasets, we generate up to two French translations for v_e, v_{f1} and v_{f2}, using the top translations from BabelNet (Navigli and Ponzetto, 2012) and Google Translate. v_{f1} and v_{f2} are then paired back with w_e, thus generating two unannotated crosslingual examples. Annotation is crowdsourced on Crowdflower [2]: for each example pair (w_e, v_f), workers are asked to label it as true ($w_e \rightarrow v_f$) or false ($w_e \not\rightarrow v_f$). We select the positive examples annotated with high-agreement, and obtain the same number of negative examples by applying the same translation process to negative examples[3].

[2] http://crowdflower.com

[3] Manual annotation is unnecessary for negative examples: it is unlikely that a negative example $w \not\rightarrow v$ will become positive by translating v automatically. We verified this by asking a bilingual speaker to manually check a random sample of 100 such translated pairs, which were all found to remain negative.

5.3 Crowdsourcing Cross-lingual Entailment Judgments

Detecting lexical entailment for bilingual word pairs is a non-trivial annotation task, and requires a good command of both French and English. For quality control, we first ask a bilingual speaker in our group to conduct a pilot annotation task, which we use to evaluate workers' ability to perform the task. In addition, Crowdflower allows us to present this task to only workers who have a proven knowledge of French, and to georestrict the task to countries most likely to have French-English bilinguals.

N	Examples
0	(*animal*,*couleur*), (*animal*,*reptile*), (*art*, *serpent*)
1	(*asp*, *vertébré*), (*chancellor*, *guide*), (*psychotherapy*,*capacité*)
2	(*bookmark*, *marque*), (*postman*, *ouvrier*), (*endurance*, *force*)
3	(*cricket*,*insecte*), (*muse*,*divinité*), (*parapet*, *paroi*)
4	(*ape*, *animal*), (*reimbursement*,*paiement*), (*lady*,*adulte*)
5	(*epistle*,*lettre)*, (*gin*,*boisson*), (*potato*,*nourriture*)

Table 3: Randomly selected examples for each level of annotator agreement: N is the number of annotators who labeled pair as true (out of five)

This approach yielded a large number of high-quality annotations quickly. 1680 cross-lingual pairs were presented to five annotators each. 24 pairs did not receive enough judgments. For the remaining 1656 pairs, four or more annotators agreed for 75% of examples (Figure 1).

This result first shows that we can indeed generate a gold standard for the challenging task of cross-lingual lexical entailment using such crowdsourcing techniques. We ensure high-quality annotations by selecting all 945 (w, v) where four or more annotators agree that $w \rightarrow v$.

In addition, the degree of agreement sheds light on how the notion of lexical entailment is interpreted by non-expert annotators. In Table 3, we present randomly selected examples for each agree-

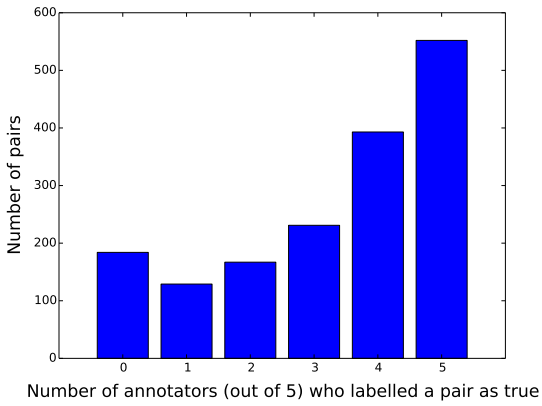

Figure 1: Agreement statistics for dataset creation. X-axis indicates number of annotators who labeled pair as true (out of five)

ment level. The bottom two rows represent clear positive examples, that are cross-lingual equivalents of hypernymy or synonymy relations: e.g., *gin* is a kind of drink (*boisson* in French). The top row represent clear negative examples, where the two words are unrelated (e.g., *art* and *serpent*, which means *snake* in English) or the directionality is wrong (e.g., *animal* ← *reptile*). The middle rows where one to three annotators chose to annotate the word pair as negative contain less clear-cut cases, including association relations (e.g., *endurance* → *force*), and examples where entailment judgments requires taking into account more subtle word sense or translation distinctions (e.g., *bookmark* can be translated as *marque* for a positive example, but the most frequent sense of *marque* translates into English as *brand*, for which the entailment relation does not hold.)

6 Experimental Conditions

We estimate the following models for evaluation on the test sets described in the previous section.

6.1 Sparse Bilingual Model

Estimating sparse, bilingual representations as described in Section 4.2 first requires learning dense monolingual representations in two languages ($\mathbf{X_e}$ and $\mathbf{X_f}$). We can in principle use any type of dense representations. We choose to train GloVe (Pennington et al., 2014) vectors on a corpus comprised of Gigaword and Wikipedia to learn dense representations of 2000 dimensions for English and French. English vectors are trained on a corpus of 4.9B words, while French vectors are trained on 1.2B words.

Second, we construct $\mathbf{S}$ by word-aligning large amounts of parallel corpora using a fast implementation of IBM model 2 (Dyer et al., 2013). We combine Europarl (Koehn, 2005), News Commentary [4], and Wikipedia [5] to create a large parallel corpus of 72M English tokens and 78M French tokens. All corpora are tokenized and lowercased.

We learn 100-dimensional sparse representations with hyperparameters $\lambda_e = \lambda_f = 0.5, \lambda_x = 10$.

6.2 Contrastive Models

We also learn two other sets of 100-dimensional word representations, as a basis for comparison.

First, we learn sparse monolingual English word representations, which will be used in monolingual lexical entailment experiments (Section 7.1). These are trained using the non-negative sparse method described in Section 4.1, on the same 4.9B word English corpus that was used for learning bilingual representations.

Second, we learn dense bilingual word representations using BiCVM (Hermann and Blunsom, 2014), to use as a baseline for our cross-lingual lexical entailment experiments (Section 7.2). BiCVM uses sentence aligned parallel corpora to learn representations for words in two languages, with the objective that when these representations are composed into representations for parallel sentences, the Euclidean distance between the parallel sentences should be minimized. We learn English-French vectors on the parallel corpora described in Section 6.1.

7 Results

7.1 Monolingual Tasks

We first evaluate the monolingual version of our sparse model on English test sets. While our focus is on the cross-lingual setting, the monolingual evaluation lets us compare a version of our newly proposed approach with existing lexical entailment results (Levy et al., 2015), obtained using dense word representations compared with cosine similarity. This is not a controlled comparison, as training conditions are not comparable. Nevertheless it

[4] http://www.statmt.org/wmt15/training-parallel-nc-v10.tgz

[5] https://sites.google.com/site/iwsltevaluation2015/data-provided

English Dataset	Levy et al.	Sparse+cosine	Sparse+*balAPinc*
Baroni et al. (2012)	.788	.745	.744
Baroni and Lenci (2011)	.197	.552	.546
Kotlerman et al. (2010)	.461	.620	.618
Turney and Mohammad (2015)	.642	.576	.587

Table 4: Evaluating sparse representations on monolingual lexical entailment (F-score): we compare previously published unsupervised results (Levy et al.) to our sparse word representations. While this is not a controlled comparison, we can see that our word representations yield roughly comparable performance to prior work.

is reassuring to see that sparse word representations are roughly on par with previously published results. This suggests that they indeed provide good features for discovering entailment relations, using both cosine and *balAPinc* as metrics[6].

Results (Table 4) show that sparse representations lead to performance comparable to previous approaches, thus providing a strong motivation for using the same for the crosslingual task.

7.2 Cross-lingual Task

Word Representations	Cosine	*balAPinc*
bilingual + dense	.528	.548
monolingual + sparse	.663	.675
bilingual + sparse	.687	**.703**

Table 5: F-Score on Cross-lingual Lexical Entailment Task. All results are obtained by 10-fold cross-validation. Using *balAPinc* with features from the sparse bilingual representations outperforms all other approaches.

We evaluate our proposed approach on the new English-French lexical entailment test set. We evaluate the impact of choosing a sparse representation by comparing our approach to the dense bilingual word representations obtained with the BiCVM model (Section 6.2). We also evaluate the usefulness of bilingual vs. monolingual word representations: given a bilingual example (w_e, v_f), we translate v_f into English using Google Translate, and then detect lexical entailment using English sparse representations for the English pair (w_e, v_e) as described in Section 6.2.

[6]While cosine and *balAPinc* yield comparable F-scores here, *balAPinc* is still a better metric as it captures directionality. If the test sets included examples of both entailment direction for every pair, cosine would yield incorrect predictions for as many as half of the examples, since its predictions would be the same regardless of the direction.

Results are summarized in Table 5. First, we observe that *balAPinc* outperforms cosine for all word representations, confirming that the directional metric is better suited to discovering lexical entailment. Second, all sparse models significantly outperform the model based on dense representation, which suggests that sparsity helps discover useful context features. Finally, our proposed approach (*balAPinc* with features from sparse bilingual representations) yields the best result overall, perofrming better than the second best model (cosine with features from sparse bilingual representations) by approximately 1.6 points. This difference is highly statistically significant (at $p < 0.01$) according to the McNemar's Test (Dietterich, 1998). Our model also outperforms translation followed by monolingual entailment, confirming the need for models that directly compare the meaning of words across languages, instead of using translation as a proxy.

8 Discussion

8.1 Examining bilingual dimensions learned

One motivation for using sparse representations is that they yield interpretable dimensions: one can summarize a dimension using the top scoring words in its column. Interpreting five randomly selected dimensions learned in our bilingual model (Table 6) shows that we indeed learn English and French dimensions that align well, but that are not identical - reflecting the difference in contexts observed in monolingual English vs. French corpora, as needed to detect lexical entailment.

8.2 Sparse Vectors Help Capture Distributional Inclusion

One advantage of our sparse representations over dense bilingual representations is that they can better leverage an asymmetric scoring function like

French Dimensions	English Dimensions
logiciel, fichiers, web, microsoft	files, web, microsoft, www
université, collège, lycée, conseil de administration	university, college, graduate, faculty
virus informatique, virus, infection, cellules	virus, viruses, infection, cells
doigts, genoux, jambes, muscles	bruises, fingers, toes, knees
budapest, stockholm, copenhague, buenos	lahore, dhaka, harare, karachi

Table 6: Top scoring words in 5 randomly selected French and English dimensions learned by our bilingual model.

balAPinc . Consider the following two pairs from our dataset - *(mesothelioma,tumeur)* and *(tumor,mésothéliome)*. The former is a positive example since *mesothelioma* → *tumeur*, but the latter is negative (since not all tumors are mesotheliomas.)

Cosine similarity is unable to differentiate between these two cases, assigning a high score to both these pairs, causing both of them to be labeled positive. However, *balAPinc* with sparse representations teases them apart by giving a high score to the first pair and a low score to the second.

In the bilingual sparse model, *mesothelioma* and *mésothéliome* have only one non-zero entry (in the dimension corresponding to ['virus', 'viruses', 'infection', 'cells', 'cancer']) whereas *tumeur* and *tumor* have five non-zero entries in their representations. Based on the distributional inclusion hypothesis, this difference in the number of non-zero entries is a strong basis for *mesothelioma* → *tumor*.

8.3 Benefits of Bilingual Modeling

Examining the results of the approach based on translation followed by monolingual entailment confirms the problems raised by sense ambiguities.

Consider the English word *drug*, which can be translated into the French *drogue* when used in the narcotics sense, and *médicament* when used in the medicinal sense. Thus the pair *(antibiotic,drogue)* that is correctly labeled as negative in the crosslingual case, gets converted to *(antibiotic,drug)* by translation and is then incorrectly labeled as positive. Similarly, the pair (*coriander, herbe*), which is positive in the crosslingual case, gets translated to (*coriander, grass*) because the French *herbe* is primarily aligned to the English *grass* (rather than *herb*). The translated pair is labeled negative.

9 Related Work

Bilingual Word Representations Much recent work targets the problem of learning low-dimensional multilingual word representations, using matrix decomposition techniques such as Principal Component Analysis and Canonical Correlation Analysis (Gaussier et al., 2004; Jagarlamudi and Daumé III, 2012; Gardner et al., 2015), Latent Dirichlet Allocation (Mimno et al., 2009; Jagarlamudi and Daumé III, 2010), and neural distributional representations (Klementiev et al., 2012; Gouws et al., 2015; Lu et al., 2015, among others). However, these models have typically been evaluated on translation induction or document categorization, which, unlike lexical entailment, focus on capturing coarse cross-lingual correspondences.

Sparse Word Representations While cooccurrence matrices and their PPMI transformed variants are early examples of sparse representations, recent work has leveraged Nonnegative Sparse Embedding (NNSE) (Murphy et al., 2012). These models have been augmented to incorporate different types of linguistically motivated constraints, such as compositionality of words into phrases (Fyshe et al., 2015), or a hierarchical regularizer that captures knowledge of word relations (Yogatama et al., 2015).

Sparse representations have also been used for monolingual lexical entailment in the Boolean Distributional Semantic Model (Kruszewski et al., 2015), which shares our hypothesis on the usefulness of sparsity in meaning representations. However, they are meant to be used in different settings: while the boolean features can interestingly capture formal semantics, they are not as useful in our unsupervised setting, since they do not provide the feature rankings required to use the *balAPinc* metric.

Cross-Lingual Semantic Analysis To the best of our knowledge, lexical entailment has not been pre-

viously addressed in a cross-lingual setting. The long tradition of lexical semantic analysis in cross-lingual settings has mostly focused on using translations to characterize word meaning (Diab and Resnik, 2002; Carpuat and Wu, 2007; Lefever and Hoste, 2010; McCarthy et al., 2013, among others). An exception is Cross-lingual *Textual* Entailment (Mehdad et al., 2010), which aims to detect whether an English hypothesis H entails a text T written in another language. We plan to use our lexical models to address this task in the future.

10 Conclusion

In this work, we introduced the task of cross-lingual lexical entailment, which aims to detect whether the meaning of a word in one language can be inferred from the meaning of a word in another language. We constructed a dataset with gold annotations through crowdsourcing, and presented a top-performing solution based on novel sparse bilingual word representations that leverages both word co-occurrence patterns in monolingual corpora and bilingual correspondences learned in parallel text[7].

A key limitation of this work is that we address lexical entailment out of context, based on word representations that collapse multiple word senses into a single vector . These could be addressed in future work by adapting existing methods for learning sense-specific representations for dense vectors (Jauhar et al., 2015; Ettinger et al., 2016; Reisinger and Mooney, 2010; Guo et al., 2014; Huang et al., 2012; Neelakantan et al., 2015) to our sparse representations, and target cross-lingual textual entailment tasks, which focus on full sentences rather than isolated words. We also plan to study lexical entailment on more languages and example types, as well as investigate the usefulness of our bilingual representations in higher level multilingual applications such as machine translation.

Acknowledgments

The authors would like to thank Tom Goldstein, Roberto Navigli and Peter Turney for their assistance with tools and datasets, Philip Resnik and the CLIP lab at the University of Maryland for stimulating discussions, and the reviewers for their insightful feedback. This work was partially funded by an Amazon Academic Research Award.

[7]Data and code are available at `http://cs.umd.edu/~yogarshi`.

References

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *EACL 2012*, pages 23–32. Association for Computational Linguistics.

Luisa Bentivogli and Emanuelle Pianta. 2000. Looking for lexical gaps. In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000: Stuttgart, Germany, August 8th-12th, 2000*, pages 663–669.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, June.

Mona Diab and Philip Resnik. 2002. An Unsupervised Method for Word Sense Tagging using Parallel Text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia, Pennsylvania, July.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics.

Allyson Ettinger, Philip Resnik, and Marine Carpuat. 2016. Retrofitting sense-specific word vectors using parallel text. In *Proceedings of NAACL*.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *ACL 2015*, pages 1491–1500.

Christiane Fellbaum and Piek Vossen. 2012. Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2):313–326.

Alona Fyshe, Leila Wehbe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2015. A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 32–41, Denver, Colorado, May–June. Association for Computational Linguistics.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Matt Gardner, Kejun Huang, Evangelos Papalexakis, Xiao Fu, Partha Talukdar, Christos Faloutsos, Nicholas Sidiropoulos, and Tom Mitchell. 2015. Translation invariant word embeddings. In *EMNLP 2015*, pages 1084–1088.

E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte, and H. Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maayan Geffet and Ido Dagan. 2005a. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114. Association for Computational Linguistics.

Maayan Geffet and Ido Dagan. 2005b. The distributional inclusion hypotheses and lexical entailment. In *ACL 2005*.

Tom Goldstein, Christoph Studer, and Richard Baraniuk. 2014. A field guide to forward-backward splitting with a FASTA implementation. *arXiv eprint*, abs/1411.3406.

Tom Goldstein, Christoph Studer, and Richard Baraniuk. 2015. FASTA: A generalized implementation of forward-backward splitting, January. http://arxiv.org/abs/1501.04979.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING*, pages 497–507.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *ACL*.

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: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32Nd European Conference on Advances in Information Retrieval*, ECIR'2010, pages 444–456, Berlin, Heidelberg. Springer-Verlag.

Jagadeesh Jagarlamudi and Hal Daumé III. 2012. Regularized interlingual projections: Evaluation on multilingual transliteration. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 12–23, Jeju Island, Korea, July. Association for Computational Linguistics.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of NAACL*.

John Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP 2007*, pages 967–975.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India, December. The COLING 2012 Organizing Committee.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2009. Directional distributional similarity for lexical expansion. In *ACL-IJCNLP 2009*, pages 69–72. Association for Computational Linguistics.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04):359–389.

German Kruszewski, Denis Paperno, and Marco Baroni. 2015. Deriving boolean structures from distributional vectors. *Transactions of the Association for Computational Linguistics*, 3:375–388.

Els Lefever and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden, July.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *NAACL HLT 2015*, pages 970–976.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, Denver, Colorado, May–June. Association for Computational Linguistics.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM.

Diana McCarthy, Ravi Som Sinha, and Rada Mihalcea. 2013. The cross-lingual lexical substitution task. *Language Resources and Evaluation*, 47(3):607–638.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards cross-lingual textual entailment. In *NAACL 2010*, pages 321–324.

David Mimno, Hanna M. 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: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING 2012*, pages 1933–1950, Mumbai, India, December. The COLING 2012 Organizing Committee.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1512–1522, Beijing, China, July. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP 2014*, pages 1532–1543.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Diana Santos. 1990. Lexical gaps and idioms in machine translation. In *Proceedings of the 13th conference on Computational linguistics-Volume 2*, pages 330–335.

Peter D Turney and Saif M Mohammad. 2015. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 21(03):437–476.

Dani Yogatama, Manaal Faruqui, Chris Dyer, and Noah Smith. 2015. Learning word representations with hierarchical sparse coding. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 87–96.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461.