Word Sense Disambiguation with Multilingual Features

Carmen Banea and Rada Mihalcea
Department of Computer Science and Engineering
University of North Texas
carmenbanea@my.unt.edu, rada@cs.unt.edu

Abstract

This paper explores the role played by a multilingual feature representation for the task of word sense disambiguation. We translate the context of an ambiguous word in multiple languages, and show through experiments on standard datasets that by using a multilingual vector space we can obtain error rate reductions of up to 25%, as compared to a monolingual classifier.

1 Introduction

Ambiguity is inherent to human language. In particular, word sense ambiguity is prevalent in all natural languages, with a large number of the words in any given language carrying more than one meaning. For instance, the English noun *plant* can mean *green plant* or *factory*; similarly the French word *feuille* can mean *leaf* or *paper*. The correct sense of an ambiguous word can be selected based on the context where it occurs, and correspondingly the problem of *word sense disambiguation* is defined as the task of automatically assigning the most appropriate meaning to a polysemous word within a given context.

Among the various knowledge-based (Lesk, 1986; Mihalcea et al., 2004) and data-driven (Yarowsky, 1995; Ng and Lee, 1996) word sense disambiguation methods that have been proposed to date, supervised systems have been constantly observed as leading to the highest performance. In these systems, the sense disambiguation problem is formulated as a supervised learning task, where each sense-tagged occurrence of a particular word is transformed into a feature vector which is then used in an automatic learning process. One of the main drawbacks associated with these methods is the fact that their performance is closely connected to the amount of labeled data available at hand.

In this paper, we investigate a new supervised word sense disambiguation method that is able to take additional advantage of the sense-labeled examples by exploiting the information that can be obtained from a multilingual representation. We show that by representing the features in a multilingual space, we are able to improve the performance of a word sense disambiguation system by a significant margin, as compared to a traditional system that uses only monolingual features.

2 Related Work

Despite the large number of word sense disambiguation methods that have been proposed so far, targeting the resolution of word ambiguity in different languages, there are only a few methods that try to explore more than one language at a time. The work that is perhaps most closely related to ours is the bilingual bootstrapping method introduced in (Li and Li, 2002), where word translations are automatically disambiguated using information iteratively drawn from two languages. Unlike that approach, which iterates between two languages to select the correct translation for a given target word, in our method we *simultaneously* use the features extracted from several languages. In fact, our method can handle more than two languages at a time, and we show that the accuracy of the disambiguation algorithm increases with the number of languages used.

There have also been a number of attempts to exploit parallel corpora for word sense disambiguation (Resnik and Yarowsky, 1999; Diab and Resnik, 2002; Ng et al., 2003), but in that line of work the parallel texts were mainly used as a way to induce word senses or to create sense-tagged corpora, rather than as a source of additional multilingual views for the disambiguation features. Another related technique is concerned with the selection of correct word senses in context using large corpora in a second language (Dagan and Itai, 1994), but as before, the additional language is used to help distinguishing between the word senses in the original language, and not as a source of additional information for the disambiguation context.

Also related is the recent SEMEVAL task that has been proposed for cross-lingual lexical substitution, where the word sense disambiguation task was more flexibly formulated as the identification of cross-lingual lexical substitutes in context (Mihalcea et al., 2010). A number of different approaches have been proposed by the teams participating in the task, and although several of them involved the translation of contexts or substitutes from one language to another, none of them attempted to make simultaneous use of the information available in the two languages.

Finally, although the multilingual subjectivity classifier proposed in Banea et al. (2010) is not directly applicable to the disambiguation task we address in this paper, their findings are similar to ours. In that paper, the authors showed how a natural language task can benefit from the use of features drawn from multiple languages, thus supporting the hypothesis that multilingual features can be effectively used to improve the accuracy of a monolingual classifier.

3 Motivation

Our work seeks to explore the expansion of a monolingual feature set with features drawn from multiple languages in order to generate a more robust and more effective vector-space representation that can be used for the task of word sense disambiguation. While traditional monolingual representations allow a supervised learning systems to achieve a certain accuracy, we try to surpass this limitation by infusing additional information in the model, mainly in the form of features extracted from the machine translated view of the monolingual data. A statistical machine translation (MT) engine does not only provide a dictionary-based translation of the words surrounding a given ambiguous word, but it also encodes the translation knowledge derived from very large parallel corpora, thus accounting for the contextual dependencies between the words.

In order to better explain why a multilingual vector space provides for a better representation for the word sense disambiguation task, consider the following examples centered around the ambiguous verb **build**.[1] For illustration purposes, we only show examples for four out of the ten possible meanings in WordNet (Fellbaum, 1998), and we only show the translations in one language (French). All the translations are performed using the Google Translate engine.

> En 1: Telegraph Co. said it will spend $20 million to **build** a factory in Guadalajara, Mexico, to make telephone answering machines. (*sense id 1*)
> Fr 1: Telegraph Co. a annoncé qu'il dépensera 20 millions de dollars pour **construire** une usine á Guadalajara, au Mexique, pour faire répondeurs téléphoniques.
>
> En 2: A member in the House leadership and skilled legislator, Mr. Fazio nonetheless found himself burdened not only by California's needs but by Hurricane Hugo amendments he accepted in a vain effort to **build** support in the panel. (*sense id 3*)
> Fr 2: Un membre de la direction de la Chambre et le législateur compétent, M. Fazio a néanmoins conclu lui-même souffre, non seulement par les besoins de la Californie, mais par l'ouragan Hugo amendements qu'il a accepté dans un vain effort pour **renforcer** le soutien dans le panneau.
>
> En 3: Burmah Oil PLC, a British independent oil and specialty-chemicals marketing concern, said SHV Holdings N.V. has **built** up a 7.5% stake in the company. (*sense id 3*)

[1] The sentences provided and their annotations are extracted from the SEMEVAL corpus.

> Fr 3: Burmah Oil PLC, une huile indépendant britannique et le souci de commercialisation des produits chimiques de spécialité, a déclaré SHV Holdings NV a **acquis** une participation de 7,5% dans la société.
>
> En 4: Plaintiffs' lawyers say that buildings become "sick" when inadequate fresh air and poor ventilation systems lead pollutants to **build** up inside. (*sense id 2*)
> Fr 4: Avocats des plaignants disent que les bâtiments tombent malades quand l'insuffisance d'air frais et des systèmes de ventilation insuffisante de plomb polluants de s'**accumuler** à l'intérieur.

As illustrated by these examples, the multilingual representation helps in two important ways. First, it attempts to disambiguate the target ambiguous word by assigning it a different translation depending on the context where it occurs. For instance, the first example includes a usage for the verb **build** in its most frequent sense, namely that of **construct** (WordNet: *make by combining materials and parts*), and this sense is correctly translated into French as **construire**. In the second sentence, **build** is used as part of the verbal expression **build support** where it means *to form or accumulate steadily* (WordNet), and it is accurately translated in both French sentences as **renforcer**. For sentences three and four, **build** is followed by the adverb **up**, yet in the first case, its sense id in WordNet is 3, *build or establish something abstract*, while in the second one is 2, *form or accumulate steadily*. Being able to infer from the co-occurrence of additional words appearing the context, the MT engine differentiates the two usages in French, translating the first occurrence as **acquis** and the second one as **accumuler**.

Second, the multilingual representation also significantly enriches the feature space, by adding features drawn from multiple languages. For instance, the feature vector for the first example will not only include English features such as *factory* and *make*, but it will also include additional French features such as *usine* and *faire*. Similarly, the second example will have a feature vector including words such as *buildings* and *systems*, and also *bâtiments* and *systèmes*. While this multilingual representation can sometime result in redundancy when there is a one-to-one translation between languages, in most cases however the translations will enrich the feature space, by either indicating that two features in English share the same meaning (e.g., the words *manufactory* and *factory* will both be translated as *usine* in French), or by disambiguating ambiguous English features using different translations (e.g., the context word *plant* will be translated in French as *usine* or *plante*, depending on its meaning).

Appending therefore multilingual features to the monolingual vector generates a more orthogonal vector space. If, previously, the different senses of **build** were completely dependent on their surrounding context in the source language, now they are additionally dependent on the disambiguated translation of **build** given its context, as well as the context itself and the translation of the context.

4 Multilingual Vector Space Representations for WSD

4.1 Datasets

We test our model on two publicly available word sense disambiguation datasets. Each dataset includes a number of ambiguous words. For each word, a number of sample contexts were extracted and then manually labeled with their correct sense. Therefore, both datasets follow a Zipfian distribution of senses in context, given their natural usage. Note also that senses do not cross part-of-speech boundaries.

The TWA[2] (two-way ambiguities) dataset contains sense tagged examples for six words that have two-way ambiguities (bass, crane, motion, palm, plant, tank). These are words that have been previously used in word sense disambiguation experiments reported in (Yarowsky, 1995; Mihalcea, 2003). Each word has approximately 100 to 200 examples extracted from the British National Corpus. Since the words included in this dataset have only two homonym senses, the classification task is easier.

[2] http://www.cse.unt.edu/~rada/downloads.html\#twa

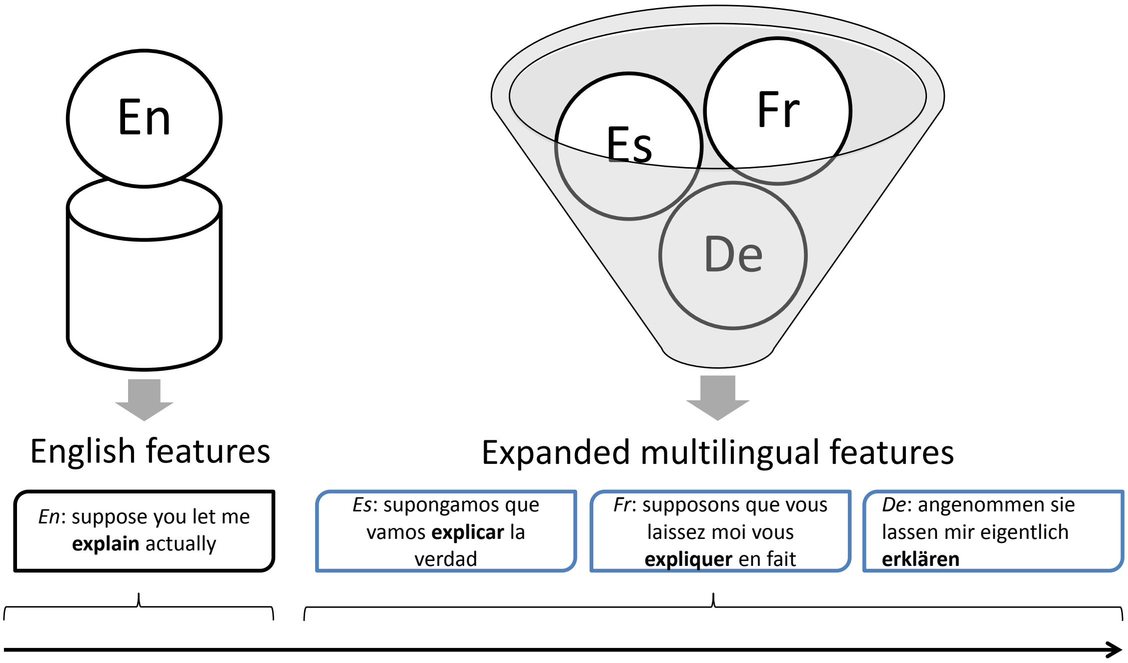
En
Es
Fr
De
English features
Expanded multilingual features
En: suppose you let me **explain** actually
Es: supongamos que vamos **explicar** la verdad
Fr: supposons que vous laissez moi vous **expliquer** en fait
De: angenommen sie lassen mir eigentlich **erklären**

Figure 1: Construction of a multilingual vector (combinations of target languages $C(3, k)$, where $k = 0..3$

The second dataset is the SEMEVAL corpus 2007 (Pradhan et al., 2007),[3] consisting of a sample of 35 nouns and 65 verbs with usage examples extracted from the Penn Treebank as well as the Brown corpus, and annotated with OntoNotes sense tags (Hovy et al., 2006). These senses are more coarse grained when compared to the traditional sense repository encoded in the WordNet lexical database. While OntoNotes attains over 90% inter-annotator agreement, rendering it particularly useful for supervised learning approaches, WordNet is too fine grained even for human judges to agree (Hovy et al., 2006). The number of examples available per word and per sense varies greatly; some words have as few as 50 examples, while some others can have as many as 2,000 examples. Some of these contexts are considerably longer than those appearing in TWA, containing around 200 words. For the experiments reported in this paper, given the limitations imposed by the number of contexts that can be translated by the online translation engine,[4] we randomly selected a subset of 31 nouns and verbs from this dataset.

4.2 Model

In order to generate a multilingual representation for the TWA and SEMEVAL datasets, we rely on the method proposed in Banea et al. (2010) and use Google Translate to transfer the data from English into several other languages and produce multilingual representations. We experiment with three languages, namely French (Fr), German (De) and Spanish (Es). Our choice is motivated by the fact that when Google made public their statistical machine translation system in 2007, these were the only languages covered by their service, and we therefore assume that the underlying statistical translation models are also the most robust. Upon translation, the data is aligned at instance level, so that the original English context is augmented with three mirroring contexts in French, German, and Spanish, respectively.

We extract the word unigrams from each of these contexts, and then generate vectors that consist of the original English unigrams followed by the multilingual portion resulted from all possible combinations of the three languages taken 0 through 3 at a time, or more formally $C(3, k)$, where $k = 0..3$ (see Figure 1). For instance, a vector resulting from $C(3, 0)$ is the traditional monolingual vector, whereas a vector built from the combination $C(3, 3)$ contains features extracted from all languages.

[3] `http://nlp.cs.swarthmore.edu/semeval/tasks/task17/description.shtml`

[4] We use Google Translate (`http://translate.google.com/`), which has a limitation of 1,000 translations per day.

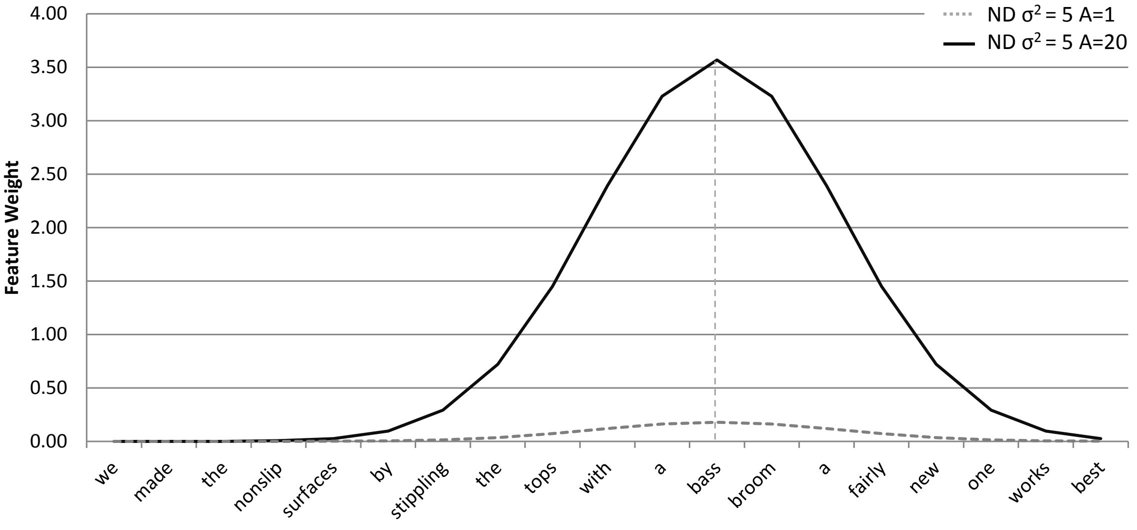

Figure 2: Example of sentence whose words are weighted based on a normal distribution with variance of 5, and an amplification factor of 20

4.2.1 Feature Weighting

For weighting, we use a parametrized weighting based on a normal distribution scheme, to better leverage the multilingual features. Let us consider the following sentence:

> We made the non-slip surfaces by stippling the tops with a <head> bass </head> broom a fairly new one works best.

Every instance in our datasets contains an XML-marking before and after the word to be disambiguated (also known as a headword), in order to identify it from the context. For instance, in the example above, the headword is **bass**. The position of this headword in the context can be considered the mean of a normal distribution. When considering a $\sigma^2 = 5$, five words to the left and right of the mean are activated with a value above 10^{-2} (see the dotted line in Figure 2). However, all the features are actually activated by some amount, allowing this weighting model to capture a continuous weight distribution across the entire context. In order to attain a higher level of discrepancy between the weight of consecutive words, we amplify the normal distribution curve by an empirically determined factor of 20, effectively mapping the values to an interval ranging from 0 to 4. We apply this amplified activation to every occurrence of a headword in a context. If two activation curves overlap, meaning that a given word has two possible weights, the final weight is set to the highest (generated by the closest headword in context). Similar weighting is also performed on the translated contexts, allowing for the highest weight to be attained by the headword translated into the target language, and a decrementally lower weight for its surrounding context.

This method therefore allows the vector-space model to capture information pertaining to both the headword and its translations in the other languages, as well as a language dependent gradient of the neighboring context usage. While a traditional bigram or trigram model only captures an exact expression, a normal distribution based model is able to account for wild cards, and transforms the traditionally sparse feature space into one that is richer and more compact at the same time.

4.3 Adjustments

We encountered several technical difficulties in translating the XML-formatted datasets, which we will expand on in this section.

4.3.1 XML-formatting and alignment

First of all, as every instance in our datasets contains an XML-marked headword (as shown in Section 4.2.1), the tags interfere with the MT system, and we had to remove them from the context before proceeding with the translation. The difficulty came from the fact that the translated context provides no way of identifying the translation of the original headword. In order to acquire candidate translations of the English headword we query the Google Multilingual Dictionary[5] (setting the dictionary direction from English to the target language) and consider only the candidates listed under the correct part-of-speech. We then scan the translated context for any of the occurrences mined from the dictionary, and locate the candidates.

In some of the cases we also identify candidate headwords in the translated context that do not mirror the occurrence of a headword in the English context (i.e., the number of candidates is higher than the number of headwords in English). We solve this problem by relying on the assumption that there is an ideal position for a headword candidate, and this ideal position should reflect the relative position of the original headword with regard to its context. This alignment procedure is supported by the fact that the languages we use follow a somewhat similar sentence structure; given parallel paragraphs of text, these cross-lingual "context anchors" will lie in close vicinity. We therefore create two lists: the first list is the reference English list, and contains the indexes of the English headwords (normalized to 100); the second list contains the normalized indexes of the candidate headwords in the target language context. For each candidate headword in the target language, we calculate the shortest distance to a headword appearing in the reference English list. Once the overall shortest distance is found, both the candidate headword's index in the target language and its corresponding English headword's index are removed from their respective list. The process continues until the reference English list is empty.

4.3.2 Inflections

There are also cases when we are not able to identify a headword due to the fact that we are trying to find the lemma (extracted from the multilingual dictionary) in a fully inflected context, where most probably the candidate translation is inflected as well. As French, German and Spanish are all highly inflected languages, we are faced with two options: to either lemmatize the contexts in each of the languages, which requires a lemmatizer tuned for each language individually, or to stem them. We chose the latter option, and used the Lingua::Stem::Snowball,[6] which is a publicly available implementation of the Porter stemmer in multiple languages.

To summarize, all the translations are stemmed to obtain maximum coverage, and alignment is performed when the number of candidate entries found in a translated context does not match the frequency of candidate headwords in the reference English context. Also, all the contexts are processed to remove any special symbols and numbers.

5 Results and Discussion

5.1 Experimental Setup

In order to determine the effect of the multilingual expanded feature space on word sense disambiguation, we conduct several experiments using the TWA and SEMEVAL datasets. The results are shown in Tables 1 and 2.

Our proposed model relies on a multilingual vector space, where each individual feature is weighted using a scheme based on a modified normal distribution (Section 4.2.1). As eight possible combinations are available when selecting one main language (English) and combinations of three additional languages

[5] `http://www.google.com/dictionary`

[6] `http://search.cpan.org/dist/Lingua-Stem-Snowball/lib/Lingua/Stem/Snowball.pm`

taken 0 through 3 at a time (Spanish, French and German), we train eight Naïve Bayes learners[7] on the resulted datasets: one monolingual (En), three bilingual (En-De, En-Fr, En-Es), three tri-lingual (En-De-Es, En-De-Fr, En-Fr-Es), and one quadri-lingual (En-Fr-De-Es). Each dataset is evaluated using ten fold cross-validation; the resulting micro-accuracy measures are averaged across each of the language groupings and they appear in Tables 1 and 2 in ND-L1 (column 4), ND-L2 (column 5), ND-L3 (column 6), and ND-L4 (column 7), respectively. Our hypothesis is that as more languages are added to the mix (and therefore the number of features increases), the learner will be able to distinguish better between the various senses.

5.2 Baselines

Our baseline consists of the predictions made by a majority class learner, which labels all examples with the predominant sense encountered in the training data.[8] Note that the most frequent sense baseline is often times difficult to surpass because many of the words exhibit a disproportionate usage of their main sense (i.e., higher than 90%), such as the noun *bass* or the verb *approve*. Despite the fact that the majority vote learner provides us with a supervised baseline, it does not take into consideration actual features pertaining to the instances. We therefore introduce a second, more informed baseline that relies on binary-weighted features extracted from the English view of the datasets and we train a multinomial Naïve Bayes learner on this data. For every word included in our datasets, the binary-weighted Naïve Bayes learner achieves the same or higher accuracy as the most frequent sense baseline.

5.3 Experiments

Comparing the accuracies obtained when training on the monolingual data, the binary weighted baseline surpasses the normal distribution-based weighting model in only three out of six cases on the TWA dataset (difference ranging from .5% to 4.81%), and in 6 out of 31 cases on the SEMEVAL dataset (difference ranging from .53% to 7.57%, where for 5 of the words, the difference is lower than 3%). The normal distribution-based model is thus able to activate regions around a particular headword, and not an entire context, ensuring more accurate sense boundaries, and allowing this behavior to be expressed in multilingual vector spaces as well (as seen in columns 7-9 in Tables 1 and 2).

When comparing the normal distribution-based model using one language versus more languages, 5 out of 6 words in TWA score highest when the expanded feature space includes all languages, and one scores highest for combinations of 3 languages (only .17% higher than the accuracy obtained for all languages). We notice the same behavior in the SEMEVAL dataset, where 18 of the words exhibit their highest accuracy when all four languages are taken into consideration, and 3 achieve the highest score for three-language groupings (at most .37% higher than the accuracy obtained for the four language grouping). While the model displays a steady improvement as more languages are added to the mix, four of the SEMEVAL words are unable to benefit from this expansion, namely the verbs buy (-0.61%), care (-1.45%), feel (-0.29%) and propose (-2.94%). Even so, we are able to achieve error rate reductions ranging from 6.52% to 63.41% for TWA, and from 3.33% to 34.62% for SEMEVAL.

To summarize the performance of the model based on the expanded feature set and the proposed baselines, we aggregate all the accuracies from Tables 1 and 2, and present the results obtained in Table 3. The monolingual modified normal-distribution model is able to exceed the most common sense baseline and the binary-weighted Naïve Bayes learner for both datasets, proving its superiority as compared to a purely binary-weighted model. Furthermore, we notice a consistent increase in accuracy as more languages are added to the vector space, displaying an average increment of 1.7% at every step for TWA, and 0.67% for SEMEVAL. The highest accuracy is achieved when all languages are taken into consideration: 86.02% for TWA and 83.36% for SEMEVAL, corresponding to an error reduction of 25.96% and 10.58%, respectively.

[7] We use the multinomial Naïve Bayes implementation provided by the Weka machine learning software (Hall et al., 2009).

[8] Our baseline it is not the same as the traditional most common sense baseline that uses WordNet's first sense heuristic, because our data sets are not annotated with WordNet senses.

1 Word	2 # Inst	3 # Senses	4 MCS	5 BIN-L1	6 ND-L1	7 ND-L2	8 ND-L3	9 ND-L4	10 Error Red.
bass.n	107	2	90.65	90.65	90.65	91.28	91.90	**92.52**	20.00
crane.n	95	2	75.79	75.79	76.84	76.14	76.49	**78.95**	9.09
motion.n	201	2	70.65	81.09	79.60	86.73	89.88	**92.54**	63.41
palm.n	201	2	71.14	73.13	87.06	88.89	**89.72**	89.55	19.23
plant.n	187	2	54.55	79.14	74.33	77.90	81.82	**83.96**	37.50
tank.n	201	2	62.69	77.61	77.11	76.29	76.45	**78.61**	6.52

Table 1: Accuracies obtained on the TWA dataset; Columns: **1** - words contained in the corpus, **2** - number of examples for a given word, **3** - number of senses covered by the examples, **4** - micro-accuracy obtained when using the most common sense (MCS), **5** - micro-accuracy obtained using the multinomial Naïve Bayes classifier on binary weighted monolingual features in English, **6** - **9** - average micro-accuracy computed over all possible combinations of English and 3 languages taken 0 through 3 at a time, resulted from features weighted following a modified normal distribution with $\sigma^2 = 5$ and an amplification factor of 20 using a multinomial Naïve Bayes learner, where **6** - one language, **7** - 2 languages, **8** - 3 languages, **9** - 4 languages, **10** - error reduction calculated between ND-L1 (6) and ND-L4 (9)

6 Conclusion

This paper explored the cumulative ability of features originating from multiple languages to improve on the monolingual word sense disambiguation task. We showed that a multilingual model is suited to better leverage two aspects of the semantics of text by using a machine translation engine. First, the various senses of a target word may be translated into other languages by using different words, which constitute unique, yet highly salient features that effectively expand the target word's space. Second, the translated context words themselves embed co-occurrence information that a translation engine gathers from very large parallel corpora. This information is infused in the model and allows for thematic spaces to emerge, where features from multiple languages can be grouped together based on their semantics, leading to a more effective context representation for word sense disambiguation. The average micro-accuracy results showed a steadily increasing progression as more languages are added to the vector space. Using two standard word sense disambiguation datasets, we showed that a classifier based on a multilingual representation can lead to an error reduction ranging from 10.58% (SEMEVAL) to 25.96% (TWA) as compared to the monolingual classifier.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340 and IIS award #1018613. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Banea, C., R. Mihalcea, and J. Wiebe (2010, August). Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, pp. 28–36.

Dagan, I. and A. Itai (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics 20*(4), 563–596.

Diab, M. and P. Resnik (2002, July). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40st Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, PA.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Word	# Inst	# Senses	MCS	BIN-L1	ND-L1	ND-L2	ND-L3	ND-L4	Error Red.
approve.v	53	2	94.34	94.34	94.34	94.34	95.60	**96.23**	33.33
ask.v	348	6	64.94	68.39	72.41	73.66	74.71	**75.00**	9.37
bill.n	404	3	65.10	88.12	90.59	91.75	92.41	**92.82**	23.68
buy.v	164	5	**78.66**	**78.66**	78.05	77.64	77.44	77.44	-2.78
capital.n	278	4	92.81	92.81	92.81	92.81	93.17	**93.53**	10.00
care.v	69	3	78.26	78.26	**86.96**	86.47	85.99	85.51	-11.11
effect.n	178	3	82.02	82.02	84.83	85.96	**86.33**	85.96	7.41
exchange.n	363	5	71.90	73.83	78.51	82.37	84.85	**85.95**	34.62
explain.v	85	2	88.24	88.24	88.24	88.24	88.24	88.24	0.00
feel.v	347	3	**82.13**	**82.13**	**82.13**	82.04	81.94	81.84	-1.61
grant.v	19	2	63.16	73.68	73.68	71.93	71.93	**78.95**	20.00
hold.v	129	8	34.88	**45.74**	43.41	43.41	43.41	43.41	0.00
hour.n	187	4	**84.49**	**84.49**	83.96	83.78	83.78	**84.49**	3.33
job.n	188	3	74.47	74.47	80.32	80.67	82.62	**84.04**	18.92
part.n	481	4	81.91	81.91	82.12	83.30	84.13	**85.45**	18.60
people.n	754	4	91.11	91.11	91.11	91.29	92.22	**93.37**	25.37
point.n	469	9	71.64	73.99	77.61	82.09	83.51	**84.22**	29.52
position.n	268	7	27.61	60.82	61.19	66.17	**68.91**	68.66	19.23
power.n	251	3	47.81	**84.46**	76.89	81.94	82.87	83.27	27.59
president.n	879	3	86.23	89.87	87.14	88.28	89.34	**90.79**	28.32
promise.v	50	2	**88.00**	**88.00**	86.00	86.67	87.33	**88.00**	14.29
propose.v	34	2	85.29	85.29	**88.24**	87.25	86.27	85.29	-25.00
rate.n	1009	2	84.64	86.92	87.02	88.07	88.64	**89.30**	17.56
remember.v	121	2	99.17	99.17	99.17	99.17	99.17	99.17	0.00
rush.v	28	2	92.86	92.86	92.86	92.86	92.86	92.86	0.00
say.v	2161	5	97.78	97.78	97.78	97.78	97.78	97.78	0.00
see.v	158	6	44.94	47.47	49.37	51.05	51.69	**52.53**	6.25
state.n	617	3	83.14	83.95	85.25	85.25	85.47	**85.74**	3.30
system.n	450	5	55.56	72.44	74.00	73.85	75.26	**75.78**	6.84
value.n	335	3	89.25	89.25	89.25	89.35	89.45	**89.85**	5.56
work.v	230	7	64.78	65.65	66.96	68.26	**68.99**	68.70	5.26

Table 2: Accuracies obtained on the SEMEVAL dataset; Columns: **1** - words contained in the corpus, **2** - number of examples for a given word, **3** - number of senses covered by the examples, **4** - micro-accuracy obtained when using the most common sense (MCS), **5** - micro-accuracy obtained using the multinomial Naïve Bayes classifier on binary weighted monolingual features in English, **6** - **9** - average micro-accuracy computed over all possible combinations of English and 3 languages taken 0 through 3 at a time, resulted from features weighted following a modified normal distribution with $\sigma^2 = 5$ and an amplification factor of 20 using a multinomial Naïve Bayes learner, where **6** - one language, **7** - 2 languages, **8** - 3 languages, **9** - 4 languages, **10** - error reduction calculated between ND-L1 (6) and ND-L4 (9)

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Dataset	MCS	BIN-L1	ND-L1	ND-L2	ND-L3	ND-L4	Error Red.
TWA	70.91	79.57	80.93	82.87	84.38	86.02	25.96
SEMEVAL	75.71	80.52	81.36	82.18	82.78	83.36	10.58

Table 3: Aggregate accuracies obtained on the TWA and SEMEVAL datasets; Columns: **1** - dataset, **2** - average micro-accuracy obtained when using the most common sense (MCS), **3** - average micro-accuracy obtained using the multinomial Naïve Bayes classifier on binary weighted monolingual features in English, **4** - **7** - average micro-accuracy computed over all possible combinations of English and 3 languages taken 0 through 3 at a time, resulted from features weighted following a modified normal distribution with $\sigma^2 = 5$ and an amplification factor of 20 using a multinomial Naïve Bayes learner, where **4** - one language, **5** - 2 languages, **6** - 3 languages, **7** - 4 languages, **8** - error reduction calculated between ND-L1 (4) and ND-L4 (7)

Fellbaum, C. (1998). *WordNet, An Electronic Lexical Database*. The MIT Press.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The weka data mining software: An update. *SIGKDD Explorations 11*(1).

Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel (2006). Ontonotes: the 90In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, NAACL '06, Morristown, NJ, USA, pp. 57–60. Association for Computational Linguistics.

Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*, Toronto.

Li, C. and H. Li (2002). Word translation disambiguation using bilingual bootstrapping. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania.

Mihalcea, R. (2003, September). The role of non-ambiguous words in natural language disambiguation. In *Proceedings of the conference on Recent Advances in Natural Language Processing RANLP-2003*, Borovetz, Bulgaria.

Mihalcea, R., R. Sinha, and D. McCarthy (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the ACL Workshop on Semantic Evaluations*, Uppsala, Sweden.

Mihalcea, R., P. Tarau, and E. Figa (2004). PageRank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20st International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Ng, H. and H. Lee (1996). Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL 1996)*, Santa Cruz.

Ng, H., B. Wang, and Y. Chan (2003, July). Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan.

Pradhan, S., E. Loper, D. Dligach, and M. Palmer (2007, June). Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

Resnik, P. and D. Yarowsky (1999). Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering 5*(2), 113–134.

Yarowsky, D. (1995, June). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, Cambridge, MA.