USP*wlv* and WLV*usp*: Combining Dictionaries and Contextual Information for Cross-Lingual Lexical Substitution

Wilker Aziz
University of São Paulo
São Carlos, SP, Brazil
wilker.aziz@usp.br

Lucia Specia
University of Wolverhampton
Wolverhampton, UK
l.specia@wlv.ac.uk

Abstract

We describe two systems participating in Semeval-2010's *Cross-Lingual Lexical Substitution* task: USP*wlv* and WLV*usp*. Both systems are based on two main components: (i) a dictionary to provide a number of possible translations for each source word, and (ii) a contextual model to select the best translation according to the context where the source word occurs. These components and the way they are integrated are different in the two systems: they exploit corpus-based and linguistic resources, and supervised and unsupervised learning methods. Among the 14 participants in the subtask to identify the *best* translation, our systems were ranked 2nd and 4th in terms of recall, 3rd and 4th in terms of precision. Both systems outperformed the baselines in all subtasks according to all metrics used.

1 Introduction

The goal of the *Cross-Lingual Lexical Substitution* task in Semeval-2010 (Mihalcea et al., 2010) is to find the best (***best*** subtask) Spanish translation or the 10-best (***oot*** subtask) translations for 100 different English source words depending on their context of occurrence. Source words include nouns, adjectives, adverbs and verbs. 1,000 occurrences of such words are given along with a short context (a sentence).

This task resembles that of Word Sense Disambiguation (WSD) within Machine Translation (MT). A few approaches have recently been proposed using standard WSD features to learn models using *translations* instead of *senses* (Specia et al., 2007; Carpuat and Wu, 2007; Chan and Ng, 2007). In such approaches, the global WSD score is added as a feature to statistical MT systems, along with additional features, to help the system on its choice for the best translation of a source word or phrase.

We exploit contextual information in alternative ways to standard WSD features and supervised approaches. Our two systems - USP*wlv* and WLV *usp* - use two main components: (i) a list of possible translations for the source word regardless of its context; and (ii) a contextual model that ranks such translations for each occurrence of the source word given its context.

While these components constitute the core of most WSD systems, the way they are created and integrated in our systems differs from standard approaches. Our systems do not require a model to disambiguate / translate each particular source word, but instead use general models. We experimented with both corpus-based and standard dictionaries, and different learning methodologies to rank the candidate translations. Our main goal was to maximize the accuracy of the system in choosing the ***best*** translation.

WLV*usp* is a very simple system based essentially on (i) a Statistical Machine Translation (SMT) system trained using a large parallel corpus to generate the n-best translations for each occurrence of the source words and (ii) a standard English-Spanish dictionary to filter out noisy translations and provide additional translations in case the SMT system was not able to produce a large enough number of legitimate translations, particularly for the ***oot*** subtask.

USP*wlv* uses a dictionary built from a large parallel corpus using inter-language information theory metrics and an online-learning supervised algorithm to rank the options from the dictionary. The ranking is based on global and local contextual features, such as the mutual information between the translation and the words in the source context, which are trained using human annotation on the trial dataset.

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 117–122,
Uppsala, Sweden, 15-16 July 2010.

2 Resources

2.1 Parallel corpus

The English-Spanish part of Europarl (Koehn, 2005), a parallel corpus from the European Parliament proceedings, was used as a source of sentence level aligned data. The nearly 1.7M sentence pairs of English-Spanish translations, as provided by the Fourth Workshop on Machine Translation (WMT09[1]), sum up to approximately 48M tokens in each language. Europarl was used both to train the SMT system and to generate dictionaries based on inter-language mutual information.

2.2 Dictionaries

The dictionary used by WLV*usp* was extracted using the free online service *Word Reference*[2], which provides two dictionaries: Espasa Concise and Pocket Oxford Spanish Dictionary. Regular expressions were used to extract the content of the webpages, keeping only the translations of the words or phrasal expressions, and the outcome was manually revised. The manual revision was necessary to remove translations of long idiomatic expressions which were only defined through examples, for example, for the verb *check*: "we *checked* up and found out he was lying – hicimos averiguaciones y comprobamos que mentía". The resulting dictionary contains a number of open domain (single or multi-word) translations for each of the 100 source words. This number varies from 3 to 91, with an average of 12.87 translations per word. For example:

- **yet.r** = todavía, aún, ya, hasta ahora, sin embargo
- **paper.n** = artículo, papel, envoltorio, diario, periódico, trabajo, ponencia, examen, parte, documento, libro

Any other dictionary can in principle be used to produce the list of translations, possibly without manual intervention. More comprehensive dictionaries could result in better results, particularly those with explicit information about the frequencies of different translations. Automatic metrics based on parallel corpus to learn the dictionary can also be used, but we would expect the accuracy of the system to drop in that case.

[1] http://www.statmt.org/wmt09/translation-task.html

[2] http://www.wordreference.com/

The process to generate the corpus-based dictionary for USP*wlv* is described in Section 4.

2.3 Pre-processing techniques

The Europarl parallel corpus was tokenized and lowercased using standard tools provided by the WMT09 competition. Additionally, the sentences that were longer than 100 tokens after tokenization were discarded.

Since the task specifies that translations should be given in their basic forms, and also in order to decrease the sparsity due to the rich morphology of Spanish, the parallel corpus was lemmatized using TreeTagger (Schmid, 2006), a freely available part-of-speech (POS) tagger and lemmatizer. Two different versions of the parallel corpus were built using both lemmatized words and their POS tags:

Lemma Words are represented by their lemmatized form. In case of ambiguity, the original form was kept, in order to avoid incorrect choices. Words that could not be lemmatized were also kept as in their original form.

Lemma.pos Words are represented by their lemmatized form followed by their POS tags. POS tags representing content words are generalized into four groups: verbs, nouns, adjectives and adverbs. When the system could not identify a POS tag, a dummy tag was used.

The same techniques were used to pre-process the trial and test data.

2.4 Training samples

The trial data available for this task was used as a training set for the USP*wlv* system, which uses a supervised learning algorithm to learn the weights of a number of global features. For the 300 occurrences of 30 words in the trial data, the expected lexical substitutions were given by the task organizers, and therefore the feature weights could be optimized in a way to make the system result in good translations. These sentences were pre-processed in the same way the parallel corpus.

3 WLV*usp* system

This system is based on a combination of the Statistical Machine Translation (SMT) framework using the English-Spanish Europarl data and an English-Spanish dictionary built semi-automatically (Section 2.2). The parallel corpus

was lowercased, tokenized and lemmatized (Section 2.3) and then used to train the standard SMT system Moses (Koehn et al., 2007) and translate the trial/test sentences, producing the 1000-best translations for each input sentence.

Moses produces its own dictionary from the parallel corpus by using a word alignment tool and heuristics to build parallel phrases of up to seven source words and their corresponding target words, to which are assigned translation probabilities using frequency counts in the corpus. This methodology provides some very localized contextual information, which can help guiding the system towards choosing a correct translation. Additional contextual information is used by the language model component in Moses, which considers how likely the sentence translation is in the Spanish language (with a 5-gram language model).

Using the phrase alignment information, the translation of each occurrence of a source word is identified in the output of Moses. Since the phrase translations are learned using the Europarl corpus, some translations are very specific to that domain. Moreover, translations can be very noisy, given that the process is unsupervised. We therefore filter the translations given by Moses to keep only those also given as possible Spanish translations according to the semi-automatically built English-Spanish dictionary (Section 2.2). This is a general-domain dictionary, but it is less likely to contain noise.

For *best* results, only the top translation produced by Moses is considered. If the actual translation does not belong to the dictionary, the first translation in that dictionary is used. Although there is no information about the order of the translations in the dictionaries used, by looking at the translations provided, we believe that the first translation is in general one of the most frequent.

For *oot* results, the alternative translations provided by the 1000-best translations are considered. In cases where fewer than 10 translations are found, we extract the remaining ones from the handcrafted dictionary following their given order until 10 translations (when available) are found, without repetition.

WLV*usp* system therefore combines contextual information as provided by Moses (via its phrases and language model) and general translation information as provided by a dictionary.

4 USP*wlv* System

For each source word occurring in the context of a specific sentence, this system uses a linear combination of features to rank the options from an automatically built English-Spanish dictionary.

For the ***best*** subtask, the translation ranked first is chosen, while for the ***oot*** subtask, the 10 best ranked translations are used without repetition.

The building of the dictionary, the features used and the learning scheme are described in what follows.

Dictionary Building The dictionary building is based on the concept of inter-language Mutual Information (MI) (Raybaud et al., 2009). It consists in detecting which words in a source-language sentence trigger the appearance of other words in its target-language translation. The inter-language MI in Equation 3 can be defined for pairs of source (s) and target (t) words by observing their occurrences at the sentence level in a parallel, sentence aligned corpus. Both simple (Equation 1) and joint distributions (Equation 2) were built based on the English-Spanish Europarl corpus using its *Lemma.pos* version (Section 2.3).

$$p_l(x) = \frac{count_l(x)}{Total} \quad (1)$$

$$p_{en,es}(s,t) = \frac{f_{en,es}(s,t)}{Total} \quad (2)$$

$$MI(s,t) = p_{en,es}(s,t) log\left(\frac{p_{en,es}(s,t)}{p_{en}(s)p_{es}(t)}\right) \quad (3)$$

$$Avg_{MI}(t_j) = \frac{\sum_{i=1}^{l} w(|i-j|)MI(s_i,t_j)}{\sum_{i=1}^{l} w(|i-j|)} \quad (4)$$

In the equations, $count_l(x)$ is the number of sentences in which the word x appear in a corpus of l-language texts; $count_{en,es}(s,t)$ is the number of sentences in which source and target words co-occur in the parallel corpus; and $Total$ is the total number of sentences in the corpus of the language(s) under consideration. The distributions p_{en} and p_{es} are monolingual and can been extracted from any monolingual corpus.

To prevent discontinuities in Equation 3, we used a smoothing technique to avoid null probabilities. We assume that any monolingual event occurs at least once and the joint distribution is smoothed by a Guo's factor $\alpha = 0.1$ (Guo et al., 2004):

$$p_{en,es}(s,t) \leftarrow \frac{p_{en,es}(s,t) + \alpha p_{en}(s)p_{es}(t)}{1+\alpha}$$

For each English source word, a list of Spanish translations was produced and ranked according to inter-language MI. From the resulting list, the 50-best translations constrained by the POS of the original English word were selected.

Features The inter-language MI is a feature which indicates the global suitability of translating a source token s into a target one t. However, inter-language MI is not able to provide local contextual information, since it does not take into account the source context sentence c. The following features were defined to achieve such capability:

Weighted Average MI (aMI) consists in averaging the inter-language MI between the target word t_j and every source word s in the context sentence c (Raybaud et al., 2009). The MI component is scaled in a way that long range dependencies are considered less important, as shown in Equation 4. The scaling factor $w(\cdot)$ is assigned 1 for verbs, nouns, adjectives and adverbs up to five positions from the source word, and 0 otherwise. This feature gives an idea of how well the elements in a window centered in the source word $head$ (s_j) align to the target word t_j, representing the suitability of t_j translating s_j in the given context.

Modified Weighted Average MI (mMI) takes the average MI as previously defined, except that the source word *head* is not taken into account. In other words, the scaling function in Equation 4 equals 0 also when $|i - j| = 0$. It gives an idea of how well the source words align to the target word t_j without the strong influence of its source translation s_j. This should provide less biased information to the learning.

Best from WLV*usp* (B) consists in a flag that indicates whether a candidate t is taken as the best ranked option according to the WLV*usp* system. The goal is to exploit the information from the SMT system and handcrafted dictionary used by that system.

10-*best* from WLV*usp* (T) this feature is a flag which indicates whether a candidate t was among the 10 best ranked translations provided by the WLV*usp* system.

Online Learning In order to train a binary ranking system based on the trial dataset as our *training set*, we used the online passive-aggressive algorithm MIRA (Crammer et al., 2006). MIRA is said to be passive-aggressive because it updates the parameters only when a misprediction is detected. At training time, for each sentence a set of pairs of candidate translations is retrieved. For each of these pairs, the rank given by the system with the current parameters is compared to the correct $rank_h(\cdot)$. A loss function $loss(\cdot)$ controls the updates attributing non 0 values only for mispredictions. In our implementation, it equals 1 for any mistake made by the model.

Each element of the kind (c, s, t) = (*source context sentence*, *source head*, *translation candidate*) is assigned a feature vector $f(c, s, t) = \langle MI, aMI, mMI, B, T\rangle$, which is modeled by a vector of parameters $w \in R^5$.

The binary ranking is defined as the task of finding the best parameters w which maximize the number of successful predictions. A successful prediction happens when the system is able to rank two translation candidates as expected. For doing so, we define an oriented pair $x = (a, b)$ of candidate translations of s in the context of c and a feature vector $F(x) = f(c, s, a) - f(c, s, b)$. $signal(w \cdot F(x))$ is the orientation the model gives to x, that is, whether the system believes a is better than b or vice versa. Based on whether or not that orientation is the same as that of the reference [3], the algorithm takes the decision between updating or not the parameters. When an update occurs, it is the one that results in the minimal changes in the parameters leading to correct labeling x, that is, guaranteeing that after the update the system will rank (a, b) correctly. Algorithm 1 presents the general method, as proposed in (Crammer et al., 2006).

In the case of this binary ranking, the minimization problem has an analytic solution well defined as long as $f(c, s, a) \neq f(c, s, b)$ and $rank_h(a) \neq rank_h(b)$, otherwise $signal(w \cdot F(x))$ or the human label would not be defined, respectively. These conditions have an impact on the content of $Pairs(c)$, the set of training points built upon the system outputs for c, which can only contain pairs of differently ranked translations.

The learning scheme was initialized with a uni-

[3] Given s in the context of c and (a, b) a pair of candidate translations of s, the reference produces 1 if $rank_h(a) > rank_h(b)$ and -1 if $rank_h(b) > rank_h(a)$.

Algorithm 1 MIRA

for $c \in$ *Training Set* **do**
  for $x = (a, b) \in Pairs(c)$ **do**
    $\hat{y} \leftarrow signal(w \cdot F(x))$
    $z \leftarrow$ *correct label*(x)
    $w = argmax_u \frac{1}{2}||w - u||^2$
    s.t. $u \cdot F(x) \geq loss(\hat{y}, z)$
    $v \leftarrow v + w$
    $T \leftarrow T + 1$
  end for
end for
return $\frac{1}{T}v$

form vector. The average parameters after $N = 5$ iterations over the training set was taken.

5 Results

5.1 Official results

Tables 1 and 2 show the main results obtained by our two systems in the official competition. We contrast our systems' results against the best baseline provided by the organizers, *DIC*, which considers translations from a dictionary and frequency information from WordNet, and show the relative position of the system among the 14 participants. The metrics are defined in (Mihalcea et al., 2010).

Subtask	Metric	Baseline	WLV*usp*	Position
Best	R	24.34	25.27	4^{th}
	P	24.34	25.27	3^{rd}
	Mode R	50.34	52.81	3^{rd}
	Mode P	50.34	52.81	4^{th}
OOT	R	44.04	48.48	6^{th}
	P	44.04	48.48	6^{th}
	Mode R	73.53	77.91	5^{th}
	Mode P	73.53	77.91	5^{th}

Table 1: Official results for WLV*usp* on the test set, compared to the highest baseline, *DICT*. P = precision, R = recall. The last column shows the relative position of the system.

Subtask	Metric	Baseline	USP*wlv*	Position
Best	R	24.34	26.81	2^{nd}
	P	24.34	26.81	3^{rd}
	Mode R	50.34	58.85	1^{st}
	Mode P	50.34	58.85	2^{nd}
OOT	R	44.04	47.60	8^{th}
	P	44.04	47.60	8^{th}
	Mode R	73.53	79.84	3^{rd}
	Mode P	73.53	79.84	3^{rd}

Table 2: Official results for USP*wlv* on the test set, compared to the highest baseline, *DICT*. The last column shows the relative position of the system.

In the ***oot*** subtask, the original systems were able to output the mode translation approximately 80% of the times. From those translations, nearly 50% were actually considered as best options according to human annotators. It is worth noticing that we focused on the ***best*** subtask. Therefore, for the ***oot*** subtask we did not exploit the fact that translations could be repeated to form the set of 10 best translations. For certain source words, our resulting set of translations is smaller than 10. For example, in the WLV*usp* system, whenever the set of alternative translations identified in Moses' top 1000-best list did not contain 10 *legitimate* translations, that is, 10 translations also found in the handcrafted dictionary, we simply copied other translations from that dictionary to amount 10 different translations. If they did not sum to 10 because the list of translations in the dictionary was too short, we left the set as it was. As a result, 58% of the 1000 test cases had fewer than 10 translations, many of them with as few as two or three translations. In fact, the list of ***oot*** results for the complete test set resulted in only $1,950$ translations, when there could be $10,000$ ($1,000$ test case occurrences $*$ 10 translations). In the next section we describe some additional experiments to take this issue into account.

5.2 Additional results

After receiving the gold-standard data, we computed the scores for a number of variations of our two systems. For example, we checked whether the performance of USP*wlv* is too dependent on the handcrafted dictionary, via the features **B** and **T**. Table 3 presents the performance of two variations of USP*wlv*: MI-aMI-mMI was trained without the two contextual flag features which depend on WLV*usp*. MI-B-T was trained without the mutual information contextual features. The variation MI-aMI-mMI of USP*wlv* performs well even in the absence of the features coming from WLV*usp*, although the scores are lower. These results show the effectiveness of the learning scheme, since USP*wlv* achieves better performance by combining these feature variations, as compared to their individual performance.

To provide an intuition on the contribution of the two different components in the system WLV*usp*, we checked the proportion of times a translation was provided by each of the components. In the ***best*** subtask, 48% of the translations came from Moses, while the remaining 52% pro-

Subtask	Metric	Baseline	MI-aMI-mMI	MI-B-T
Best	R	24.34	22.59	20.50
	P	24.34	22.59	20.50
	Mode R	50.34	50.21	44.01
	Mode P	50.34	50.21	44.01
OOT	R	39.65	47.60	32.75
	P	44.04	39.65	32.75
	Mode R	73.53	74.19	56.70
	Mode P	73.53	74.19	56.70

Table 3: Comparing between variations of the system USP*wlv* on the test set and the highest baseline, *DICT*. The variations are different sources of contextual knowledge: MI (MI–aMI–mMI) and the WLV*usp* (MI–B–T) system.

vided by Moses were not found in the dictionary. In those cases, the first translation in the dictionary was used. In the ***oot*** subtask, only 12% (246) of the translations came from Moses, while the remaining $(1,704)$ came from the dictionary. This can be explained by the little variation in the n-best lists produced by Moses: most of the variations account for word-order, punctuation, etc.

Finally, we performed additional experiments in order to exploit the possibility of replicating well ranked translations for the ***oot*** subtask. Table 4 presents the results of some strategies arbitrarily chosen for such replications. For example, in the colums labelled "5" we show the scores for repeating (once) the 5 top translations. Notice that precision and recall increase as we take fewer top translation and repeat them more times. In terms of mode metrics, by reducing the number of distinct translations from 10 to 5, USP*wlv* still outperforms (marginally) the baseline. In general, the new *systems* outperform the baseline and our previous results (see Table 1 and 2) in terms of precision and recall. However, according to the other *mode* metrics, they are below our official systems.

System	Metric	*5*	*4*	*3*	*2*
WLV*usp*	R	69.09	88.36	105.32	122.29
	P	69.09	88.36	105.32	122.29
	Mode R	68.27	63.05	63.05	52.47
	Mode P	68.27	63.05	63.05	52.47
USP*wlv*	R	73.50	94.78	102.96	129.09
	P	73.50	94.78	102.96	129.09
	Mode R	73.77	68.27	62.62	57.40
	Mode P	73.77	68.27	62.62	57.40

Table 4: Comparison between different strategies for duplicating answers in the task ***oot***. The systems output a number of distinct guesses and through arbitrarily schemes replicate them in order to complete a list of 10 translations.

6 Discussion and future work

We have presented two systems combining contextual information and a pre-defined set of translations for cross-lingual lexical substitution. Both systems performed particularly well in the ***best*** subtask. A handcrafted dictionary has shown to be essential for the WLV*usp* system and also helpful for the USP*wlv* system, which uses an additional dictionary automatically build from a parallel corpus. We plan to investigate how such systems can be improved by enhancing the corpus-based resources to further minimize the dependency on the handcrafted dictionary.

References

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72.

Yee Seng Chan and Hwee Tou Ng. 2007. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics*, pages 33–40.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-agressive algorithms. *Jornal of Machine Learning Research*, 7:551–585.

Gang Guo, Chao Huang, Hui Jiang, and Ren-Hua Wang. 2004. A comparative study on various confidence measures in large vocabulary speech recognition. In *International Symposium on Chinese Spoken Language Processing*, pages 9–12.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *SemEval-2010: 5th International Workshop on Semantic Evaluations*.

Sylvain Raybaud, Caroline Lavecchia, David Langlois, and Kamel Smaili. 2009. Word- and sentence-level confidence measures for machine translation. In *13th Annual Conference of the European Association for Machine Translation*, pages 104–111.

Helmut Schmid. 2006. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Natural Language Processing*, pages 44–49.

Lucia Specia, Mark Stevenson, and Maria das Graças Volpe Nunes. 2007. Learning expressive models for word sense disambiguation. In *45th Annual Meeting of the Association for Computational Linguistics*, pages 41–148.