Low-cost Enrichment of Spanish WordNet with Automatically Translated Glosses: Combining General and Specialized Models

Jesús Giménez and **Lluís Màrquez**
TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
{jgimenez,lluism}@lsi.upc.edu

Abstract

This paper studies the enrichment of Spanish WordNet with synset glosses automatically obtained from the English WordNet glosses using a phrase-based Statistical Machine Translation system. We construct the English-Spanish translation system from a parallel corpus of proceedings of the European Parliament, and study how to adapt statistical models to the domain of dictionary definitions. We build specialized language and translation models from a small set of parallel definitions and experiment with robust manners to combine them. A statistically significant increase in performance is obtained. The best system is finally used to generate a definition for all Spanish synsets, which are currently ready for a manual revision. As a complementary issue, we analyze the impact of the amount of in-domain data needed to improve a system trained entirely on out-of-domain data.

1 Introduction

Statistical Machine Translation (SMT) is today a very promising approach. It allows to build very quickly and fully automatically Machine Translation (MT) systems, exhibiting very competitive results, only from a parallel corpus aligning sentences from the two languages involved.

In this work we approach the task of enriching Spanish WordNet with automatically translated glosses[1]. The source glosses for these translations are taken from the English WordNet (Fellbaum, 1998), which is linked, at the synset level, to Spanish WordNet. This resource is available, among other sources, through the Multilingual Central Repository (MCR) developed by the MEANING project (Atserias et al., 2004).

We start by empirically testing the performance of a previously developed English–Spanish SMT system, built from the large Europarl corpus[2] (Koehn, 2003). The first observation is that this system completely fails to translate the specific WordNet glosses, due to the large language variations in both domains (vocabulary, style, grammar, etc.). Actually, this is confirming one of the main criticisms against SMT, which is its strong domain dependence. Since parameters are estimated from a corpus in a concrete domain, the performance of the system on a different domain is often much worse. This flaw of statistical and machine learning approaches is well known and has been largely described in the NLP literature, for a variety of tasks (e.g., parsing, word sense disambiguation, and semantic role labeling).

Fortunately, we count on a small set of Spanish hand-developed glosses in MCR[3]. Thus, we move to a working scenario in which we introduce a small corpus of aligned translations from the concrete domain of WordNet glosses. This in-domain corpus could be itself used as a source for constructing a specialized SMT system. Again, experiments show that this small corpus alone does not suffice, since it does not allow to estimate good translation parameters. However, it is well suited for combination with the Europarl corpus, to generate combined Language and Translation

[1] Glosses are short dictionary definitions that accompany WordNet synsets. See examples in Tables 5 and 6.

[2] The Europarl Corpus is available at: http://-people.csail.mit.edu/people/koehn/publications/europarl

[3] About 10% of the 68,000 Spanish synsets contain a definition, generated without considering its English counterpart.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 287–294,
Sydney, July 2006.

Models. A substantial increase in performance is achieved, according to several standard MT evaluation metrics. Although moderate, this boost in performance is statistically significant according to the bootstrap resampling test described by Koehn (2004b) and applied to the BLEU metric.

The main reason behind this improvement is that the large out-of-domain corpus contributes mainly with coverage and recall and the in-domain corpus provides more precise translations. We present a qualitative error analysis to support these claims. Finally, we also address the important question of how much in-domain data is needed to be able to improve the baseline results.

Apart from the experimental findings, our study has generated a very valuable resource. Currently, we have the complete Spanish WordNet enriched with one gloss per synset, which, far from being perfect, constitutes an axcellent starting point for a posterior manual revision.

Finally, we note that the construction of a SMT system when few domain-specific data are available has been also investigated by other authors. For instance, Vogel and Tribble (2002) studied whether an SMT system for speech-to-speech translation built on top of a small parallel corpus can be improved by adding knowledge sources which are not domain specific. In this work, we look at the same problem the other way around. We study how to adapt an out-of-domain SMT system using in-domain data.

The rest of the paper is organized as follows. In Section 2 the fundamentals of SMT and the components of our MT architecture are described. The experimental setting is described in Section 3. Evaluation is carried out in Section 4. Finally, Section 5 contains error analysis and Section 6 concludes and outlines future work.

2 Background

Current state-of-the-art SMT systems are based on ideas borrowed from the Communication Theory field. Brown et al. (1988) suggested that MT can be statistically approximated to the transmission of information through a *noisy channel*. Given a sentence $f = f_1..f_n$ (distorted signal), it is possible to approximate the sentence $e = e_1..e_m$ (original signal) which produced f. We need to estimate $P(e|f)$, the probability that a translator produces f as a translation of e. By applying Bayes' rule it is decomposed into: $P(e|f) = \frac{P(f|e)*P(e)}{P(f)}$.

To obtain the string e which maximizes the translation probability for f, a search in the probability space must be performed. Because the denominator is independent of e, we can ignore it for the purpose of the search: $e = argmax_e P(f|e) * P(e)$. This last equation devises three components in a SMT system. First, a *language model* that estimates $P(e)$. Second, a *translation model* representing $P(f|e)$. Last, a *decoder* responsible for performing the arg-max search. Language models are typically estimated from large monolingual corpora, translation models are built out from parallel corpora, and decoders usually perform approximate search, e.g., by using dynamic programming and beam search.

However, in word-based models the modeling of the context in which the words occur is very weak. This problem is significantly alleviated by phrase-based models (Och, 2002), which represent nowadays the state-of-the-art in SMT.

2.1 System Construction

Fortunately, there is a number of freely available tools to build a phrase-based SMT system. We used only standard components and techniques for our basic system, which are all described below.

The *SRI Language Modeling Toolkit* (SRILM) (Stolcke, 2002) supports creation and evaluation of a variety of language models. We build trigram language models applying linear interpolation and Kneser-Ney discounting for smoothing.

In order to build phrase-based translation models, a phrase extraction must be performed on a word-aligned parallel corpus. We used the GIZA++ SMT Toolkit[4] (Och and Ney, 2003) to generate word alignments We applied the phrase-extract algorithm, as described by Och (2002), on the Viterbi alignments output by GIZA++. We work with the union of source-to-target and target-to-source alignments, with no heuristic refinement. Phrases up to length five are considered. Also, phrase pairs appearing only once are discarded, and phrase pairs in which the source/target phrase was more than three times longer than the target/source phrase are ignored. Finally, phrase pairs are scored by relative frequency. Note that no smoothing is performed.

Regarding the arg-max search, we used the *Pharaoh* beam search decoder (Koehn, 2004a), which naturally fits with the previous tools.

[4] http://www.fjoch.com/GIZA++.html

3 Data Sets and Evaluation Metrics

As a general source of English–Spanish parallel text, we used a collection of 730,740 parallel sentences extracted from the Europarl corpus. These correspond exactly to the training data from the Shared Task 2: *Exploiting Parallel Texts for Statistical Machine Translation* from the ACL-2005 Workshop on *Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*[5].

To be used as specialized source, we extracted, from the MCR , the set of 6,519 English–Spanish parallel glosses corresponding to the already defined synsets in Spanish WordNet. These definitions corresponded to 5,698 nouns, 87 verbs, and 734 adjectives. Examples and parenthesized texts were removed. Parallel glosses were tokenized and case lowered. We discarded some of these parallel glosses based on the difference in length between the source and the target. The gloss average length for the resulting 5,843 glosses was 8.25 words for English and 8.13 for Spanish. Finally, gloss pairs were randomly split into training (4,843), development (500) and test (500) sets.

Additionally, we counted on two large monolingual Spanish electronic dictionaries, consisting of 142,892 definitions (2,112,592 tokens) ('D1') (Martí, 1996) and 168,779 definitons (1,553,674 tokens) ('D2') (Vox, 1990), respectively.

Regarding evaluation, we used up to four different metrics with the aim of showing whether the improvements attained are consistent or not. We have computed the BLEU score (accumulated up to 4-grams) (Papineni et al., 2001), the NIST score (accumulated up to 5-grams) (Doddington, 2002), the General Text Matching (GTM) F-measure ($e = 1, 2$) (Melamed et al., 2003), and the METEOR measure (Banerjee and Lavie, 2005). These metrics work at the lexical level by rewarding n-gram matches between the candidate translation and a set of human references. Additionally, METEOR considers stemming, and allows for WordNet synonymy lookup.

The discussion of the significance of the results will be based on the BLEU score, for which we computed a bootstrap resampling test of significance (Koehn, 2004b).

[5] http://www.statmt.org/wpt05/.

4 Experimental Evaluation

4.1 Baseline Systems

As explained in the introduction we built two individual baseline systems. The first baseline ('EU') system is entirely based on the training data from the Europarl corpus. The second baseline system ('WNG') is entirely based on the training set from of the in-domain corpus of parallel glosses. In the second case phrase pairs occurring only once in the training corpus are not discarded due to the extremely small size of the corpus.

Table 1 shows results of the two baseline systems, both for the development and test sets. We compare the performance of the 'EU' baseline on these data sets with respect to the (in-domain) Europarl test set provided by the organizers of the ACL-2005 MT workshop. As expected, there is a very significant decrease in performance (e.g., from 0.24 to 0.08 according to BLEU) when the 'EU' baseline system is applied to the new domain. Some of this decrement is also due to a certain degree of free translation exhibited by the set of available 'quasi-parallel' glosses. We further discuss this issue in Section 5.

The results obtained by 'WNG' are also very low, though slightly better than those of 'EU'. This is a very interesting fact. Although the amount of data utilized to construct the 'WNG' baseline is 150 times smaller than the amount utilized to construct the 'EU' baseline, its performance is higher consistently according to all metrics. We interpret this result as an indicator that models estimated from in-domain data provide higher precision.

We also compare the results to those of a commercial system such as the on-line version 5.0 of SYSTRAN[6], a general-purpose MT system based on manually-defined lexical and syntactic transfer rules. The performance of the baseline systems is significantly worse than SYSTRAN's on both development and test sets. This means that a rule-based system like SYSTRAN is more robust than the SMT-based systems. The difference against the specialized 'WNG' also suggests that the amount of data used to train the 'WNG' baseline is clearly insufficient.

4.2 Combining Sources: Language Models

In order to improve results, in first place we turned our eyes to language modeling. In addition to

[6] http://www.systransoft.com/.

system	BLEU.n4	NIST.n5	GTM.e1	GTM.e2	METEOR
development					
EU-baseline	0.0737	2.8832	0.3131	0.2216	0.2881
WNG-baseline	0.1149	3.3492	0.3604	0.2605	0.3288
SYSTRAN	0.1625	3.9467	0.4257	0.2971	0.4394
test					
EU-baseline	0.0790	2.8896	0.3131	0.2262	0.2920
WNG-baseline	0.0951	3.1307	0.3471	0.2510	0.3219
SYSTRAN	0.1463	3.7873	0.4085	0.2921	0.4295
acl05-test					
EU-baseline	0.2381	6.5848	0.5699	0.2429	0.5153

Table 1: MT Results on development and test sets, for the two baseline systems compared to SYSTRAN and to the 'EU' baseline system on the ACL-2005 SMT workshop test set extracted from the Europarl Corpus. BLEU.n4 shows the accumulated BLEU score for 4-grams. NIST.n5 shows the accumulated NIST score for 5-grams. GTM.e1 and GTM.e2 show the GTM F_1-measure for different values of the e parameter ($e = 1$, $e = 2$, respectively). METEOR reflects the METEOR score.

the language model built from the Europarl corpus ('EU') and the specialized language model based on the small training set of parallel glosses ('WNG'), two specialized language models, based on the two large monolingual Spanish electronic dictionaries ('D1' and 'D2') were used. We tried several configurations. In all cases, language models are combined with equal probability. See results, for the development set, in Table 2.

As expected, the closer the language model is to the target domain, the better results. Observe how results using language models 'D1' and 'D2' outperform results using 'EU'. Note also that best results are in all cases consistently attained by using the 'WNG' language model. This means that language models estimated from small sets of in-domain data are helpful. A second conclusion is that a significant gain is obtained by incrementally adding (in-domain) specialized language models to the baselines, according to all metrics but BLEU for which no combination seems to significantly outperform the 'WNG' baseline alone. Observe that best results are obtained, except in the case of BLEU, by the system using 'EU' as translation model and 'WNG' as language model. We interpret this result as an indicator that translation models estimated from out-of-domain data are helpful because they provide recall. A third interesting point is that adding an out-of-domain language model ('EU') does not seem to help, at least combined with equal probability than in-domain models. Same conclusions hold for the test set, too.

4.3 Tuning the System

Adjusting the *Pharaoh* parameters that control the importance of the different probabilities that govern the search may yield significant improvements. In our case, it is specially important to properly adjust the contribution of the language models. We adjusted parameters by means of a software based on the *Downhill Simplex Method in Multidimensions* (William H. Press and Flannery, 2002). The tuning was based on the improvement attained in BLEU score over the development set. We tuned 6 parameters: 4 language models (λ_{lmEU}, λ_{lmD1}, λ_{lmD2}, λ_{lmWNG}), the translation model (λ_ϕ), and the word penalty (λ_w)[7].

Results improve substantially. See Table 3. Best results are still attained using the 'EU' translation model. Interestingly, as suggested by Table 2, the weight of language models is concentrated on the 'WNG' language model ($\lambda_{lmWNG} = 0.95$).

4.4 Combining Sources: Translation Models

In this section we study the possibility of combining out-of-domain and in-domain translation models aiming at achieving a good balance between precision and recall that yields better MT results.

Two different strategies have been tried. In a first stragegy we simply concatenate the out-of-domain corpus ('EU') and the in-domain corpus ('WNG'). Then, we construct the translatation model ('EUWNG') as detailed in Section 2.1. A second manner to proceed is to linearly combine the two different translation models into a single translation model ('EU+WNG'). In this case, we can assign different weights (ω) to the contribution of the different models to the search. We can also determine a certain threshold θ which allows us

[7] Final values when using the 'EU' translation model are $\lambda_{lmEU} = 0.22$, $\lambda_{lmD1} = 0$, $\lambda_{lmD2} = 0.01$, $\lambda_{lmWNG} = 0.95$, $\lambda_\phi = 1$, and $\lambda_w = -2.97$, while when using the 'WNG' translation model final values are $\lambda_{lmEU} = 0.17$, $\lambda_{lmD1} = 0.07$, $\lambda_{lmD2} = 0.13$, $\lambda_{lmWNG} = 1$, $\lambda_\phi = 0.95$, and $\lambda_w = -2.64$.

Translation Model	Language Model	BLEU.n4	NIST.n5	GTM.e1	GTM.e2	METEOR
EU	EU	0.0737	2.8832	0.3131	0.2216	0.2881
EU	WNG	0.1062	3.4831	0.3714	0.2631	0.3377
EU	D1	0.0959	3.2570	0.3461	0.2503	0.3158
EU	D2	0.0896	3.2518	0.3497	0.2482	0.3163
EU	D1 + D2	0.0993	3.3773	0.3585	0.2579	0.3244
EU	EU + D1 + D2	0.0960	3.2851	0.3472	0.2499	0.3160
EU	D1 + D2 + WNG	0.1094	3.4954	0.3690	0.2662	0.3372
EU	EU + D1 + D2 + WNG	0.1080	3.4248	0.3638	0.2614	0.3321
WNG	EU	0.0743	2.8864	0.3128	0.2202	0.2689
WNG	WNG	0.1149	3.3492	0.3604	0.2605	0.3288
WNG	D1	0.0926	3.1544	0.3404	0.2418	0.3050
WNG	D2	0.0845	3.0295	0.3256	0.2326	0.2883
WNG	D1 + D2	0.0917	3.1185	0.3331	0.2394	0.2995
WNG	EU + D1 + D2	0.0856	3.0361	0.3221	0.2312	0.2847
WNG	D1 + D2 + WNG	0.0980	3.2238	0.3462	0.2479	0.3117
WNG	EU + D1 + D2 + WNG	0.0890	3.0974	0.3309	0.2373	0.2941

Table 2: MT Results on development set, for several translation/language model configurations. 'EU' and 'WNG' refer to the models estimated from the Europarl corpus and the training set of parallel WordNet glosses, respectively. 'D1', and 'D2' denote the specialized language models estimated from the two dictionaries.

Translation Model	Language Model	BLEU.n4	NIST.n5	GTM.e1	GTM.e2	METEOR
development						
EU	EU + D1 + D2 + WNG	0.1272	3.6094	0.3856	0.2727	0.3695
WNG	EU + D1 + D2 + WNG	0.1269	3.3740	0.3688	0.2676	0.3452
test						
EU	EU + D1 + D2 + WNG	0.1133	3.4180	0.3720	0.2650	0.3644
WNG	EU + D1 + D2 + WNG	0.1015	3.1084	0.3525	0.2552	0.3343

Table 3: MT Results on development and test sets after tuning for the 'EU + D1 + D2 + WNG' language model configuration for the two translation models, 'EU' and 'WNG'.

to discard phrase pairs under a certain probability. These weights and thresholds were adjusted[8] as detailed in Subsection 4.3. Interestingly, at combination time the importance of the 'WNG' translation model ($\omega_{tmWNG} = 0.9$) is much higher than that of the 'EU' translation model ($\omega_{tmEU} = 0.1$).

Table 4 shows results for the two strategies. As expected, the 'EU+WNG' strategy consistently obtains the best results according to all metrics both on the development and test sets, since it allows to better adjust the relative importance of each translation model. However, both techniques achieve a very competitive performance. Results improve, according to BLEU, from 0.13 to 0.16, and from 0.11 to 0.14, for the development and test sets, respectively.

We measured the statistical signficance of the overall improvement in BLEU.n4 attained with respect to the baseline results by applying the bootstrap resampling technique described by Koehn (2004b). The 95% confidence intervals extracted from the test set after 10,000 samples are the following: $I_{\mathrm{EU-base}} = [0.0642, 0.0939]$, $I_{\mathrm{WNG-base}} = [0.0788, 0.1112]$, $I_{\mathrm{EU+WNG-best}} = [0.1221, 0.1572]$. Since the intervals are not ovelapping, we can conclude that the performance of the best combined method is statistically higher than the ones of the two baseline systems.

[8]We used values $\omega_{tmEU} = 0.1$, $\omega_{tmWNG} = 0.9$, $\theta_{tmEU} = 0.1$, and $\theta_{tmWNG} = 0.01$

4.5 How much in-domain data is needed?

In principle, the more in-domain data we have the better, but these may be difficult or expensive to collect. Thus, a very interesting issue in the context of our work is how much in-domain data is needed in order to improve results attained using out-of-domain data alone. To answer this question we focus on the 'EU+WNG' strategy and analyze the impact on performance (BLEU.n4) of specialized models extracted from an incrementally bigger number of example glosses. The results are presented in the plot of Figure 1. We compute three variants separately, by considering the use of the in-domain data: only for the translation model (TM), only for the language model (LM), and simultaneously in both models (TM+LM). In order

Translation Model	Language Model	BLEU.n4	NIST.n5	GTM.e1	GTM.e2	METEOR
development						
EUWNG	WNG	0.1288	3.7677	0.3949	0.2832	0.3711
EUWNG	EU + D1 + D2 + WNG	0.1182	3.6034	0.3835	0.2759	0.3552
EUWNG	EU + D1 + D2 + WNG (TUNED)	0.1554	3.8925	0.4081	0.2944	0.3998
EU+WNG	WNG	0.1384	3.9743	0.4096	0.2936	0.3804
EU+WNG	EU + D1 + D2 + WNG	0.1235	3.7652	0.3911	0.2801	0.3606
EU+WNG	EU + D1 + D2 + WNG (TUNED)	0.1618	4.1415	0.4234	0.3029	0.4130
test						
EUWNG	WNG	0.1123	3.6777	0.3829	0.2771	0.3595
EUWNG	EU + D1 + D2 + WNG	0.1183	3.5819	0.3737	0.2772	0.3518
EUWNG	EU + D1 + D2 + WNG (TUNED)	0.1290	3.6478	0.3920	0.2810	0.3885
EU+WNG	WNG	0.1227	3.8970	0.3997	0.2872	0.3723
EU+WNG	EU + D1 + D2 + WNG	0.1199	3.7353	0.3846	0.2812	0.3583
EU+WNG	EU + D1 + D2 + WNG (TUNED)	0.1400	3.8930	0.4084	0.2907	0.3963

Table 4: MT Results on development and test sets for the two strategies for combining translations models.

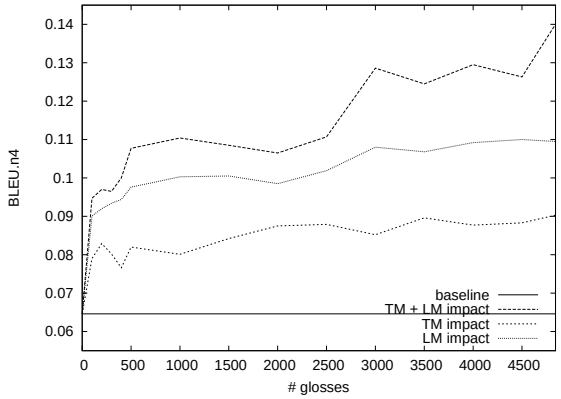

Figure 1: Impact of the size of in-domain data on MT system performance for the test set.

to avoid the possible effect of over-fitting we focus on the behavior on the test set. Note that the optimization of parameters is performed at each point in the x-axis using only the development set.

A significant initial gain of around 0.3 BLEU points is observed when adding as few as 100 glosses. In all cases, it is not until around 1,000 glosses are added that the 'EU+WNG' system stabilizes. After that, results continue improving as more in-domain data are added. We observe a very significant increase by just adding around 3,000 glosses. Another interesting observation is the boosting effect of the combination of TM and LM specialized models. While individual curves for TM and LM tend to be more stable with more than 4,000 added examples, the TM+LM curve still shows a steep increase in this last part.

5 Error Analysis

We inspected results at the sentence level based on the GTM F-measure ($e = 1$) for the best configuration of the 'EU+WNG' system. 196 sentences out from the 500 obtain an F-measure equal to or higher than 0.5 on the development set (181 sentences in the case of test set), whereas only 54 sentences obtain a score lower than 0.1. These numbers give a first idea of the relative usefulness of our system. Table 5 shows some translation cases selected for discussion. For instance, Case 1 is a clear example of unfair low score. The problem is that source and reference are not parallel but 'quasi-parallel'. Both glosses define the same concept but in a different way. Thus, metrics based on rewarding lexical similarities are not well suited for these cases. Cases 2, 3, 4 are examples of proper cooperation between 'EU' and 'WNG' models. 'EU' models provides recall, for instance by suggesting translation candidates for 'bombs' or 'price below'. 'WNG' models provide precision, for instance by choosing the right translation for 'an attack' or 'the act of'.

We also compared the 'EU+WNG' system to SYSTRAN. In the case of SYSTRAN 167 sentences obtain a score equal to or higher than 0.5 whereas 79 sentences obtain a score lower than 0.1. These numbers are slightly under the performance of the 'EU+WNG' system. Table 6 shows some translation cases selected for discussion. Case 1 is again an example of both systems obtaining very low scores because of 'quasi-parallelism'. Cases 2 and 3 are examples of SYSTRAN outperforming our system. In case 2 SYSTRAN exhibits higher precision in the translation of 'accompanying' and 'illustration', whereas in case 3 it shows higher recall by suggesting appropriate translation candidates for 'fibers', 'silkworm', 'cocoon', 'threads', and 'knitting'. Cases

F_E	F_W	F_{EW}	Source	Out_E	Out_W	Out_{EW}	Reference
0.0000	0.1333	0.1111	of the younger of two boys with the same family name	de acuerdo con el más joven de dos boys con la misma familia fama	de la younger de dos boys tiene el mismo nombre familia	de acuerdo con el más joven de dos muchachos tiene el mismo nombre familia	**que tiene menos edad**
0.2857	0.2500	0.5000	an attack by dropping bombs	atacar por cayendo **bombas**	**ataque** realizado por dropping bombs	**ataque** realizado por cayendo **bombas**	ataque con bombas
0.1250	0.7059	0.5882	the act of informing by verbal report	acto de la información por verbales ponencia	**acción y efecto** de informing por verbal **explicación**	**acción y efecto** de informaba por verbales **explicación**	acción y efecto de informar con una explicación verbal
0.5000	0.0000	0.5000	a price below the standard price	un **precio por debajo de la** norma precio	una price below número estándar price	un **precio por debajo de** la estándar precio	precio que está por debajo de lo normal

Table 5: MT output analysis of the 'EU', 'WNG' and 'EU+WNG' systems. F_E, F_W and F_{EW} refer to the GTM ($e = 1$) F-measure attained by the 'EU', 'WNG' and 'EU+WNG' systems, respectively. 'Source', Out_E, Out_W and Out_{EW} refer to the input and the output of the systems. 'Reference' corresponds to the expected output.

4 and 5 are examples where our system outperforms SYSTRAN. In case 4, our system provides higher recall by suggesting an adequate translation for 'top of something'. In case 5, our system shows higher precision by selecting a better translation for 'rate'. However, we observed that SYSTRAN tends in most cases to construct sentences exhibiting a higher degree of grammaticality.

6 Conclusions

In this work, we have enriched every synset in Spanish WordNet with a preliminary gloss, which can be later updated in a lighter process of manual revision. Though imperfect, this material constitutes a very valuable resource. For instance, WordNet glosses have been used in the past to generate sense tagged corpora (Mihalcea and Moldovan, 1999), or as external knowledge for Question Answering systems (Hovy et al., 2001).

We have also shown the importance of using a small set of in-domain parallel sentences in order to adapt a phrase-based general SMT system to a new domain. In particular, we have worked on specialized language and translation models and on their combination with general models in order to achieve a proper balance between precision (specialized in-domain models) and recall (general out-of-domain models). A substantial increase is consistently obtained according to standard MT evaluation metrics, which has been shown to be statistically significant in the case of BLEU. Broadly speaking, we have shown that around 3,000 glosses (very short sentence fragments) suffice in this domain to obtain a significant improvement. Besides, all the methods used are language independent, assumed the availability of the required in-domain additional resources.

In the future we plan to work on domain independent translation models built from WordNet itself. We may use the WordNet topology to provide translation candidates weighted according to the given domain. Moreover, we are experimenting the applicability of current Word Sense Disambiguation (WSD) technology to MT. We could favor those translation candidates showing a closer semantic relation to the source. We believe that coarse-grained is sufficient for the purpose of MT.

Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology (ALIADO TIC2002-04447-C02) and the Spanish Ministry of Education and Science (TRANGRAM, TIN2004-07925-C03-02). Our research group, TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government. Authors are grateful to Patrik Lambert for providing us with the implementation of the Simplex Method, and specially to German Rigau for motivating in its origin all this work.

References

Jordi Atserias, Luis Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek

F_{EW}	F_S	Source	Out_{EW}	Out_S	Reference
0.0000	0.0000	a newspaper that is published every day	periódico que se publica diario	un periódico que se publica cada día	publicación periódica monotemática
0.1818	0.8333	brief description accompanying an illustration	breve descripción adjuntas un aclaración	breve descripción **que acompaña una ilustración**	pequeña descripción que acompaña una ilustración
0.1905	0.7333	fibers from silkworm cocoons provide threads for knitting	fibers desde silkworm cocoons proporcionan threads para knitting	las **fibras** de los **capullos** del **gusano de seda proporcionan** los **hilos** de rosca para **hacer punto**	fibras de los capullos de gusano de seda que proporcionan hilos para tejer
1.0000	0.0000	the top of something	parte superior de una cosa	la tapa algo	parte superior de una cosa
0.6667	0.3077	a rate at which something happens	un **ritmo** al que sucede algo	una tarifa en la cual algo sucede	ritmo al que sucede una cosa

Table 6: MT output analysis of the 'EU+WNG' and SYSTRAN systems. F_{EW} and F_S refer to the GTM ($e = 1$) F-measure attained by the 'EU+WNG' and SYSTRAN systems, respectively. 'Source', Out_{EW} and Out_S refer to the input and the output of the systems. 'Reference' corresponds to the expected output.

Vossen. 2004. The MEANING Multilingual Central Repository. In *Proceedings of 2nd GWC*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Robert L. Mercer, , and Paul S. Roossin. 1988. A statistical approach to language translation. In *Proceedings of COLING'88*.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd Internation Conference on Human Language Technology*, pages 138–145.

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

Eduard Hovy, Ulf Hermjakob, and Chin-Yew Lin. 2001. The Use of External Knowledge of Factoid QA. In *Proceedings of TREC*.

Philipp Koehn. 2003. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Technical report, http://people.csail.mit.edu/-people/koehn/publications/europarl/.

Philipp Koehn. 2004a. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA'04*.

Philipp Koehn. 2004b. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP'04*.

María Antonia Martí, editor. 1996. *Gran diccionario de la Lengua Española*. Larousse Planeta, Barcelona.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of HLT/NAACL'03*.

Rada Mihalcea and Dan Moldovan. 1999. An Automatic Method for Generating Sense Tagged Corpora. In *Proceedings of AAAI*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, IBM Research Report, RC22176. Technical report, IBM T.J. Watson Research Center.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP'02*.

Stephan Vogel and Alicia Tribble. 2002. Improving Statistical Machine Translation for a Speech-to-Speech Translation Task. In *Proceedings of ICSLP-2002 Workshop on Speech-to-Speech Translation*.

Vox, editor. 1990. *Diccionario Actual de la Lengua Española*. Bibliograf, Barcelona.

William T. Vetterling William H. Press, Saul A. Teukolsky and Brian P. Flannery. 2002. *Numerical Recipes in C++: the Art of Scientific Computing*. Cambridge University Press.