Automatic Translation of WordNet Glosses

Jesús Giménez and **Lluís Màrquez**
TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
{jgimenez,lluism}@lsi.upc.edu

German Rigau
IXA Group
University of the Basque Country
rigau@si.ehu.es

Abstract

We approach the task of automatically translating the glosses in the English WordNet. We intend to generate a preliminary material which could be utilized to enrich other wordnets lacking of glosses. A Phrase-based Statistical Machine Translation system has been built using a parallel corpus of proceedings of the European Parliament. We study how to adapt the system to the domain of dictionary definitions. First, we work with specialized language models. Second, we exploit the *Multilingual Central Repository* to build domain independent translation models. Combining these two complementary techniques and properly tuning the system, a relative improvement of 64% in BLEU score is attained.

1 Introduction

In this work we study the possibility of applying Statistical Machine Translation (SMT) techniques to the glosses in the English WordNet (Fellbaum, 1998). WordNet glosses are a very useful resource. For instance, Mihalcea and Moldovan (1999) suggested an automatic method for generating sense tagged corpora which uses WordNet glosses. Hovy et al. (2001) used WordNet glosses as external knowledge to improve their Webclopedia Question Answering (QA) system.

However, most of the wordnets in the *Multilingual Central Repository* (MCR) (Atserias et al., 2004) contain very few glosses. For instance, in the current version of the Spanish WordNet fewer than 10% of the synsets have a gloss. Conversely, since version 1.6 every synset in the English WordNet has a gloss. We believe that a method to rapidly obtain glosses for all wordnets in the MCR may be helpful. These glosses could serve as a starting point for a further step of revision and post-editing. Furthermore, from a conceptual point of view, the idea of enriching the MCR using the MCR itself results very attractive.

Moreover, SMT is today a very promising approach to Machine Translation (MT) for a number of reasons. The most important one in the context of this work is that it allows to build very quickly an MT system, given only a parallel corpus representing the languages involved. Besides, SMT is fully automatic and results are also very competitive.

However, one of the main claims against SMT is that it is domain oriented. Since parameters are estimated from a parallel corpus in a specific domain, the performance of the system on a different domain is often much worse. In the absence of a parallel corpus of definitions, we built phrase-based[1] translation models on the Europarl[2] corpus (Koehn, 2003). However, the language of definitions is very specific and different to that of parliament proceedings. This is particularly harmful to the system recall, because many unknown words will be processed.

[1] The term 'phrase' used hereafter refers to a sequence of words not necessarily syntactically motivated.

[2] European Parliament Proceedings (1996-2003) are available for 11 European languages at http://people.csail.mit.edu/-people/koehn/publications/europarl/. We used a version of this corpus reviewed by the RWTH Aachen group.

In order to adapt the system to the new domain we study two separate lines. First, we use electronic dictionaries in order to build more adequate target language models. Second, we work with domain independent word-based translation models extracted from the MCR. Other authors have previously applied information extracted from aligned wordnets. Tufis et al. (2004b) presented a method for Word Sense Disambiguation (WSD) based on parallel corpora. They utilized the aligned wordnets in BalkaNet (Tufis et al., 2004a).

We suggest to use these models as a complement to phrase-based models. These two proposals together with a good tuning of the system parameters lead to a notable improvement of results. In our experiments, we focus on translation from English into Spanish. A relative increase of 64% in BLEU measure is achieved when limiting the use of the MCR-based model to the case of unknown words .

The rest of the paper is organized as follows. In Section 2 the fundamentals of SMT are depicted. In Section 3 we describe the components of our system. Experimental work is deployed in Section 4. Improvements are detailed in Section 5. Finally, in Section 6, current limitations of our approach are discussed, and further work is outlined.

2 Statistical Machine Translation

Current state-of-the-art SMT systems are based on ideas borrowed from the Communication Theory field (Weaver, 1955). 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 sentece $e = e_1..e_m$ (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 we decompose it:

$$P(e|f) = \frac{P(f|e) * P(e)}{P(f)} \qquad (1)$$

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:

$$e = argmax_e P(f|e) * P(e) \qquad (2)$$

Equation 2 devises three components in a SMT. First, a *language model* that estimates $P(e)$. Second, a *translation model* representing $P(f|e)$. Last, a *decoder* responsible for performing the search. See (Brown et al., 1993) for a detailed report on the mathematics of Machine Translation.

3 System Description

Fortunately, we can count on a number of freely available tools to build a SMT system.

We utilized the *SRI Language Modeling Toolkit* (SRILM) (Stolcke, 2002). It supports creation and evaluation of a variety of language model types based on N-gram statistics, as well as several related tasks, such as statistical tagging and manipulation of N-best lists and word lattices.

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[3] (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++. This algorithm takes as input a word alignment matrix and outputs a set of phrase pairs that is *consistent* with it. A phrase pair is said to be consistent with the word alignment if all the words within the source phrase are only aligned to words withing the target phrase, and viceversa.

Phrase pairs are scored by relative frequency (Equation 3). Let ph_f be a phrase in the source language (f) and ph_e a phrase in the target language (e). We define a function $count(ph_f, ph_e)$ which counts the number of times the phrase ph_f has been seen aligned to phrase ph_e in the training data. The conditional probability that ph_f maps into ph_e is estimated as:

$$score(ph_f|ph_e) = \frac{count(ph_f, ph_e)}{\sum_{ph_f} count(ph_f, ph_e)} \qquad (3)$$

No smoothing is performed.

[3]The GIZA++ SMT Toolkit may be freely downloaded at http://www.fjoch.com/GIZA++.html

For the search, we used the *Pharaoh* beam search decoder (Koehn, 2004). *Pharaoh* is an implementation of an efficent dynamic programming search algorithm with lattice generation and XML markup for external components. Performing an optimal decoding can be extremely costly because the search space is polynomial in the length of the input (Knight, 1999). For this reason, like most decoders, *Pharaoh* actually performs a suboptimal (beam) search by pruning the search space according to certain heuristics based on the translation cost.

4 Experiments

4.1 Experimental Setting

As sketched in Section 2, in order to build a SMT system we need to build a *language model*, and a *translation model*, all in a format that is convenient for the *Pharaoh decoder*.

We tokenized and case lowered the Europarl corpus. A set of 327,368 parallel segments of length between five and twenty was selected for training. The Spanish side consisted of 4,243,610 tokens, whereas the English side consisted of 4,197,836 tokens.

We built a trigram language model from the Spanish side of the Europarl corpus selection. Linear interpolation was applied for smoothing.

We used the GIZA++ default configuration. In the phrase extraction we worked with the union of source-to-target and target-to-source alignments, with no heuristic refinement. Only phrases up to length five were considered. Also, phrase pairs in which the source/target phrase was more than three times longer than the target/source phrase were ignored. Finally, phrase pairs appearing only once were discarded, too.

4.2 Data Sets

By means of the MCR we obtained a set of 6503 parallel glosses. These definitions correspond to 5684 nouns, 87 verbs, and 732 adjectives. Examples and parenthesized texts were removed. Gloss average length was 8,03 words for English and 7,83 for Spanish. Parallel glosses were tokenized and case lowered, and randomly split into development (3295 gloss pairs) and test (3208 gloss pairs) sets.

4.3 Evaluation Metrics

Three different evaluation metrics have been computed, namely the General Text Matching (GTM) F-measure ($e = 1, 2$) (Melamed et al., 2003), the BLEU score ($n = 4$) (Papineni et al., 2001), and the NIST score ($n = 5$) (Lin and Hovy, 2002). These metrics have proved to correlate well with both human adequacy and fluency. They all reward n-gram matches between the candidate translation and a set of reference translations. The larger the number of reference translations the more reliable these measures are. Unfortunately, in our case, a single reference translation is available.

BLEU has become a 'de facto' standard nowadays in MT. Therefore, we discuss our results based on the BLEU score. However, it has several deficiencies that turn it impractical for error analysis (Turian et al., 2003). First, BLEU does not have a clear interpretation. Second, BLEU is not adequate to work at the segment[4] level but only at the document level. Third, in order to punish candidate translations that are too long/short, BLEU computes a heuristically motivated word penalty factor.

In contrast, the GTM F-measure has an intuitive interpretation in the context of a bitext grid. It represents the fraction of the grid covered by aligned blocks. It also, by definition, works well at the segment level and punishes translations too divergent in length. Therefore, we also analyze individual cases based on the GTM F-measure.

In the future, we also consider the possibility of conducting very modest human evaluations.

4.4 Results

Baseline system results are showed in Table 1.

system	GTM-1	GTM-2	BLEU	NIST
EU-baseline-dev	0.3091	0.2196	0.0730	3.0953
EU-baseline-test	0.3028	0.2155	0.0657	3.0274
EU-europarl	0.5885	0.3567	0.2725	7.2477

Table 1: Preliminary MT Results on development (dev) and test (test) sets, and on a Europarl test set.

The performance of the system on the new domain is very low in comparison to the performance

[4] A segment is the minimal unit of parallel text. It is usually the size of a sentence. It can be smaller (a word, a phrase) or bigger (a couple of sentences, a paragraph), though.

on a set of 8490 unseen sentences from the European Parliament Proceedings.

We analyzed these results in deep detail based on the GTM F-measure ($e = 2$). Some cases are shown in Table 2. Only 28 glosses obtain an F_1 over 0.9. Most of them are too short, less than 5 words (e.g. 2917). 10% of the glosses (320) obtain an F_1 over 0.5. Interestingly, many of them are somehow related to the domain of politics and economy (e.g. 193, 293, 345, 362, 1414, 1674 and 1721). On the other hand, 18% of the glosses obtain an F_1 below 0.1. In many cases this is due to unknown vocabulary (e.g. 34, 508, 2263 and 2612). However, we found many translations unfairly scoring too low due to strong divergences between source and reference. We call this phenomenon *'quasi-parallelism'* (e.g. 7, 1606, and 2985).

5 Improvements

5.1 Language Modeling

The first improvement is based on building additional specialized language models. We utilized two large monolingual Spanish electronic dictionaries, consisting of 142,892 definitions (2,112,592 tokens) (Martí, 1996) and 168,779 definitons (1,553,674 tokens) (Vox, 1990), respectively.

We tried different language model configurations. See Table 3. We refer to the baseline system, which uses the Europarl language model only, as *'EU'*. In *'D1'* and *'D2'* we replaced the language model with those obtained from dictionaries D1 and D2, respectively. *'D1-D2'* combines the two dictionaries with equal probability. 'D1-D2-EU' combines all three language models with equal probability.

language model	GTM-1	GTM-2	BLEU	NIST
EU	0.3091	0.2196	0.0730	3.0953
D1	0.3361	0.2409	0.0905	3.4881
D2	0.3374	0.2419	0.0890	3.4719
D1-D2	0.3422	0.2457	0.0940	3.5515
D1-D2-EU	0.3428	0.2456	0.0949	3.5655

Table 3: MT Results on the development set for different language model configurations.

As expected, language models built out from dictionaries work much better than the one built from the Europarl corpus. Results improve still slightly further by combining the two dictionaries. A relative increase of 30% in BLEU score is reported. Adding the EU language model does not report any significant improvement.

5.2 Using the MCR

The second improvement is based on extracting domain independent translation models out from the MCR. Outer knowledge may be supplied to the *Pharaoh* decoder by annotating the input with alternative translation options via XML-markup. In the default setting we enrich all nouns, verbs, and adjectives by looking up all possible translations for all their meanings according to the MCR. For the 3295 glosses in the development set, a total of 13,335 words, corresponding to 8,089 nouns, 2,667 verbs and 2,579 adjectives respectively, were enriched. We have not worked on adverbs yet because of some problems with our lemmatizer. While in WordNet the lemma for adverbs is an adjective our lemmatizer returns an adverb.

Translation pairs are heuristically scored according to the number of senses which may lexicalize in the same manner. For instance, the English word *'bank'* as a noun is assigned nine different senses in WordNet. Four of these senses may lexicalize as the Spanish word *'banco'* (finantial institution) whereas only one sense lexicalizes as *'orilla'* (the bank of a river). The scoring heuristic accounts for this by assigning a higher score to '*(banco, bank)*'.

Let w_f, p_f be the source word and PoS, and w_e be the target word, we define a function $Scount(w_f, p_f, w_e)$ which counts the number of senses for (w_f, p_f) which may lexicalize as w_e. The scoring function is defined as:

$$score(w_f, p_f|w_e) = \frac{Scount(w_f, p_f, w_e)}{\sum_{(w_f,p_f)} Scount(w_f, p_f, w_e)} \quad (4)$$

In WordNet all word forms related to the same concept are grouped and represented by their lemma and part-of-speech (PoS). Therefore, input word forms must be lemmatized and PoS-tagged. WordNet takes care of the lemmatization step. For PoS-tagging we utilized the *SVMTool*[5] (Giménez and Màrquez, 2004). Similarly, at the output, the MCR

[5]The SVMTool may be freely downloaded at `http://www.lsi.upc.es/˜nlp/SVMTool/`.

case	synset-ili	Source	Target	Reference
			'good' translations	
193	00392749#n	the office and function of *president*	el cargo y función de presidente	cargo y función de presidente
293	00630513#n	the action of *attacking the enemy*	acción de atacar al enemigo	acción y efecto de atacar al enemigo
345	00785108#n	the act of giving hope or support to someone	la acción de dar esperanza o apoyo a alguien	acción de dar esperanza o apoyo a alguien
362	00804210#n	the combination of two or more *commercial companies*	la combinación de dos o más comerciales compañías	combinación de dos o más empresas
1414	05359169#n	the act of *presenting a proposal*	el acto de presentar una propuesta	acto de presentar una propuesta
1674	06089036#n	a *military unit* that is part of an *army*	unidad militar que forma parte de un ejército	unidad militar que forma parte de un ejército
1721	06213619#n	a group of *representatives* or *delegates*	grupo de representantes o delegados	grupo de representantes o delegados
2917	01612822#v	perform an action	realizar una acción	realizar una acción
			'bad' translations	
7	00012865#n	a feature of the mental life of a living organism	una característica de la vida mental de un organismo vivo	rasgo psicológico
34	00029442#n	the act of departing politely	el acto de *departing politely*	acción de marcharse de forma educada
508	02581431#n	a kitchen appliance for disposing of garbage	*kitchen* una *appliance* para *disposing* de *garbage*	cubo donde se depositan los residuos
1606	05961082#n	people in general	gente en general	grupo de gente que constituye la mayoría de la población y que define y mantiene la cultura popular y las tradiciones
2263	07548871#n	a painter of theatrical scenery	una *painter* de *theatrical scenery*	persona especializada en escenografía
2612	10069279#n	rowdy behavior	*rowdy behavior*	comportamiento escandaloso
2985	00490201#a	without reservation	sin reservas	movido por una devoción o un compromiso entusiasta y decidido

Table 2: MT examples of the baseline system. 'Source' and 'Target' refer to the input and output of the system, respectively. 'Reference' corresponds to the expected output.

provides us with lemmas instead of word forms as translation candidates. A lemma extension must be performed. We utilized components from the *Freeling*[6] package (Carreras et al., 2004) for this step. See an example of enriched input in Table 4.

Then, we proceeded applying the MCR-based model. Several strategies were tried. In all cases we allowed the decoder to bypass the MCR-based model when a better solution was found using the phrase-based model alone. See results in Table 5.

We defined as new baseline the system which combines the three language models as detailed in Subsection 5.1 (no-MCR). In a first attempt, we enriched all content words in the validation set with all possible translation candidates (ALL). No improvement was achieved. By inspecting input data, apart from some PoS-tagging errors, we found that the number of translation options generated via MCR was growing too fast for words with too many senses, particularly verbs. In order to reduce the degree of polysemy we tried limiting to words with 1, 2, 3, 4 and 5 different senses at most (S1, S2, S3, S4 and S5). Results improved slightly.

Ideally, one would wish to work with accurately word sense disambiguated input. We tried restricting translation candidates to those generated by the most frequent sense only (ALL-mfs). There was no significant variation in results.

[6]Freeling Suite of Language Analyzers may be downloaded at `http://www.lsi.upc.es/~nlp/freeling/`

<NN english="consecuciones|consecución|logro|logros|realizaciones|realización" prob="0.1666|0.1666|0.1666|0.1666|0.1666|0.1666">**accomplishment**</NN>**of an objective**

an organism such as an<NN english="insecto|insectos" prob="0.5|0.5">**insect**</NN>**that habitually shares the**<NN english="madriguera|madrigueras|nido|nidos" prob="0.25|0.25|0.25|0.25"> **nest**</NN>**of a species of**<NN english="hormiga|hormigas" prob="0.5|0.5">**ant**</NN>

the part of the human<NN english="pierna|piernas" prob="0.5|0.5">**leg**</NN>
between the<NN english="rodilla|rodillas" prob="0.5|0.5">**knee**</NN>
and the<NN english="tobillo|tobillos" prob="0.5|0.5">**ankle**</NN>

a<JJ english="casada|casadas|casado|casados" prob="0.25|0.25|0.25|0.25"> **married**</JJ>**man**

an<NN english="abstracciones|abstracción|extracciones|extracción|generalizaciones|generalización|pintura abstracta" prob="0.3333|0.3333|0.0666|0.0666|0.0666|0.0666|0.0666">**abstraction**</NN>**belonging to or**<JJ english="característica|características|característico|característicos|típica|típicas|típico|típicos" prob="0.125|0.125|0.125|0.125|0.125|0.125|0.125|0.125">characteristic</JJ>**of two**<NNS english="entidad|entidades" prob="0.5|0.5">**entities**</NNS>**or**<NNS english="partes" prob="1"> **parts**</NNS>**together**

strengthening the concentration by removing<JJ english="irrelevante|irrelevantes" prob="0.5|0.5">**extraneous**</JJ>**material**

Table 4: A sample of enriched input, scored as detailed in Equation 4.

strategy	GTM-1	GTM-2	BLEU	NIST
no-MCR	0.3428	0.2456	0.0949	3.5655
ALL	0.3382	0.2439	0.0949	3.4980
ALL-mfs	0.3367	0.2434	0.0951	3.4720
S1	0.3432	0.2469	0.0961	3.5774
S2	0.3424	0.2464	0.0963	3.5686
S3	0.3414	0.2459	0.0963	3.5512
S4	0.3412	0.2458	0.0966	3.5441
S5	0.3403	0.2451	0.0962	3.5286
N-mfs	0.3361	0.2428	0.0944	3.4588
V-mfs	0.3428	0.2456	0.0945	3.5649
A-mfs	0.3433	0.2462	0.0959	3.5776
UNK-mfs	0.3538	0.2535	0.1035	3.7580
UNK-and-S1	0.3463	0.2484	0.0977	3.6313
UNK-or-S1	0.3507	0.2523	0.1026	3.7104

Table 5: MT Results on the development set, using the MCR.

We also studied the behavior of the model applied separately to nouns (N-mfs), verbs (V-mfs), and adjectives (A-mfs). The system worked worst for nouns, and seemed to work a little better for adjectives than for verbs.

All in all, we did not find an adequate manner to have the two translation models, to cooperate properly. Therefore we decided to use the MCR-based model only for those words unknown[7] to the phrase-based model (UNK-mfs). A significant relative improvement of 9% in BLEU score was achieved.

Finally, we tried translating only those words that were both unknown and monosemous (UNK-and-S1), and those that were either unknown or monosemous (UNK-or-S1). Results did not improve.

5.3 Tuning the System

Another path we explored is the tuning of the *Pharaoh* parameters that control the importance of the different probabilities that govern the search.

In general, there are 4 important parameters to adjust: the language model probability (λ_{lm}), the translation model probability (λ_{ϕ}), the distortion probability (λ_d) and the word penalty factor (λ_w). Recall, for instance, the difference in length between source and target seen in Subsection 4.2. Tuning the λ_w parameter leads to better results. Also, a proper tuning of the probabilities of the three language models yields a significant improvement.

We utilized a sotware based on the *Downhill Simplex Method in Multidimensions* (William H. Press and Flannery, 2002). Parameters were tuned for the 'no-MCR' and 'UNK-mfs' strategies on the development set. A further relative gain of 9% in BLEU score is reported. See Table 6.

We analyzed results by the 'UNK-mfs' and 'ALL-

[7] 7.87% of the words in the development set are unknown.

strategy	GTM-1	GTM-2	BLEU	NIST
no-MCR-dev	0.3428	0.2456	0.0949	3.5655
UNK-mfs-dev	0.3538	0.2535	0.1035	3.7580
no-MCR-test	0.3352	0.2420	0.0915	3.4802
UNK-mfs-test	0.3478	0.2500	0.0991	3.6946
no-MCR-dev-T	0.3492	0.2496	0.1026	3.5352
UNK-mfs-dev-T	0.3599	0.2582	0.1124	3.7609
noMCR-test-T	0.3431	0.2450	0.0965	3.4628
UNK-mfs-test-T	0.3554	0.2546	0.1075	3.7079

Table 6: MT Results for the 'no-MCR' and 'UNK-mfs' strategies, before and after tuning (T) on development (dev) and test (test) sets.

mfs' strategies based on the GTM F-measure ($e = 2$). Table 7 shows some cases where MCR-based models prove their usefulness (e.g. 29, 35, 194, 268, 351, 377 and 965) and some cases where they cause the system to make a mistake (e.g. 1001, 1125 and 2570).

6 Conclusions

By working with specialized language models and MCR-based translation models we achieved a relative gain of 63.62% in BLEU score (0.0657 vs 0.1075) when porting the system to a new domain.

But there is a strong limitation in our approach. When we markup the input to Pharaoh we are somehow forcing the decoder to choose between a word-to-word translation and a phrase-to-phrase translation. In SMT phrase-based models have been demonstrated to outperform word-based ones. A better way to integrate MCR-based models with phrase-based models should be investigated.

Moreover, more sophisticated heuristics should be considered for selecting and scoring MCR-based translation candidates.

Finally, better results should be obtained by working with word sense disambiguated text. We could favor those translation candidates showing a closer semantic relation to the source. We believe that coarse-grained WSD is sufficient for the purpose of MT. In the short term, we plan to utilize the system by Castillo et al. (2004), winner in the Senseval-3 workshop shared task on WSD of WordNet glosses.

Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology (ALIADO TIC2002-04447-C02) and the European Comision (MEANING IST-2001-34460). 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 also grateful to Xavier Carreras for pointing us the availability of the Pharaoh decoder, and to Patrik Lambert for providing us with the implementation of the Simplex Method.

References

Jordi Atserias, Luis Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*, Brno, Czech Republic, January. ISBN 80-210-3302-9.

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*.

Peter E Brown, Stephen A. Della Pietra, Robert L. Mercer, and Vincent J. Della Pietra. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th LREC*.

Mauro Castillo, Francis Real, Jordi Atserias, and German Rigau. 2004. The talp systems for disambiguating wordnet glosses. In *Proceedings of SENSEVAL-3 Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Task: Word-Sense Disambiguation of WordNet Glosses*.

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

Jesús Giménez and Lluís Màrquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of 4th LREC*.

Eduard Hovy, Ulf Hermjakob, and Chin-Yew Lin. 2001. The use of external knowledge of factoid qa. In *Proceedings of TREC*.

Kevin Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4).

Philipp Koehn. 2003. Europarl: A multilingual corpus for evaluation of machine translation. Technical report, http://people.csail.mit.edu/people/koehn/-publications/europarl/.

case	synset-ili	Source	Target-base	Target-MCR	Reference
			UNK-mfs		
29	00025788#n	accomplishment of an objective	accomplishment de un objetivo	*consecución* de un objetivo	consecución de un objetivo
194	00393890#n	the position of secretary	situación de secretary	el cargo de *secretario*	posición de secretario
268	00579072#n	the activity of making portraits	actividad de hacer portraits	actividad de hacer *retratos*	actividad de hacer retratos
377	00913742#n	an organism such as an insect that habitually shares the nest of a species of ant	un organismo como un insect que habitually comparte el nest de una especie de ant	un organismo como un insecto que habitualmente comparte el *nido* de una especie de *hormiga*	organismo que comparte el nido de una especie de hormigas
965	04309478#n	the part of the human leg between the knee and the ankle	parte de la persona leg entre los knee y el ankle	parte de la persona *pierna* entre la *rodilla* y el *tobillo*	parte de la pierna humana comprendida entre la rodilla y el tobillo
			ALL-mfs		
35	00029961#n	the act of withdrawing	el acto de retirar	el acto de *retirarse*	acción de retirarse
351	00790504#n	a favorable judgment	una sentencia favorable	una *opinión* favorable	opinión favorable
1001	04395081#n	source of difficulty	fuente de dificultad	fuente de *problemas*	fuente de dificultad
1125	04634158#n	the branch of biology that studies plants	rama de la biología que estudios plantas	rama de la biología que estudia *factoría*	rama de la biología que estudia las plantas
2570	10015334#n	balance among the parts of something	equilibrio entre las partes de algo	equilibrio entre las partes de *entidades*	equilibrio entre las partes de algo

Table 7: MT examples of the 'ALL-mfs' and 'UNK-mfs' strategies. 'Source' refers to the raw input. 'Target-base' and 'Target-MCR' refer to the output of the baseline and MCR helped systems, respectively. 'Reference' corresponds to the expected output.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA'04*.

Chin-Yew Lin and E.H. Hovy. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Technical report, National Institute of Standards and Technology.

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, Germany.

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*.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004a. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal on Science Technology of Information. Special Issue on Balkanet*, 7(3-4):9–44.

Dan Tufis, Radu Ion, and Nancy Ide. 2004b. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of COLING'04*.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of MT SUMMIT IX*.

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

Warren Weaver. 1955. *Translation*. Machine Translation of Languages. MIT Press, Cambridge, MA.

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.