Statistical Machine Translation of WordNet glosses

Document Number	Working Paper 3.9
Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Project URL	http://www.lsi.upc.es/~nlp/meaning/meaning.html
Availability	Project Internal
Authors: Jesus Gimenez, Lluis Marquez, and German Rigau	

meaning

INFORMATION SOCIETY TECHNOLOGIES

Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Security (Distribution level)	Project Internal
Contractual date of delivery	February 2005
Actual date of delivery	February 24, 2005
Document Number	Working Paper 3.9
Type	Report
Status & version	v Draft
Number of pages	16
WP contributing to the deliberable	WP3
WPTask responsible	Bernardo Magnini
Authors	Jesus Gimenez, Lluis Marquez, and German Rigau
Other contributors	
Reviewer	
EC Project Officer	Evangelia Markidou
Authors: Jesus Gimenez, Lluis Marquez, and German Rigau	
Keywords: keys	

Abstract: State-of-the-art phrase-based Statistical Machine Translation (SMT) methods are applied to the problem of translating WordNet glosses. The system is built on top of freely available components, namely the GIZA++ SMT Toolkit, the *Pharaoh* decoder and the *SRI Language Modeling Toolkit*. Results for the English-Spanish translation task are reported.
Furthermore, novel ways to incorporate the information contained in the *Multilingual Central Repository* (MCR) into the SMT system are currently under study. Experimental results are presented.

Contents

1 Introduction **3**

2 Experimental Setting **4**

3 Results **5**

4 Using WordNet **8**

5 Discussion and Further Work **10**

1 Introduction

Nowadays, with the availability of most of the necessary components, it takes only a little effort to build a Statistical Machine Translation (SMT) system. In principle, the only additional resource required is a Parallel Corpus representing the translation process between the two languages involved.

Current state-of-the-art SMT systems are based on ideas borrowed from the Communication Theory field [Weaver, 1955]. In the late 80's [Brown *et al.*, 1988] and early 90's [Brown *et al.*, 1990] researchers at IBM T.J Watson Research Center suggested that Machine Translation (MT) could be statistically approximated to the transmission of information through a *noisy channel*[1].

Therefore, 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. To do so, 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 (Equation 1):

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

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 (Equation 2). See [Pietra *et al.*, 1993] for a detailed report on the mathematics of Machine Translation.

$$e = argmax_e P(f|e) * P(e) \tag{2}$$

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.

It is possible to build a language model out from a monolingual corpus. And it is possible to build a translation model from a parallel corpus. The decoder is a search procedure.

Fortunately, it turns out that implementations for two of these components are publically available. The *SRI Language Modeling Toolkit* (SRILM) [Stolcke, 2002] has been used to build language models. For decoding, the *Pharaoh* beam search decoder for phrase-based MT [Koehn, 2004] has been utilized.[2]

The SRILM toolkit 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.

The *Pharaoh* decoder 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

[1] This idea makes sense theoretically only if both sentences (signals) are translations of each other.

[2] Authors are grateful to Xavier Carreras for pointing us the availability of the Pharaoh decoder.

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. The beam size can be defined by threshold and histogram pruning.

Therefore, the only component we built ourselves is the translation model. However, we didn't do so from scratch. The GIZA++ SMT Toolkit has been used to generate word alignment models [Och and Ney, 2003]. A phrase extraction has been performed on the output of GIZA++ as suggested by [Och, 2002], to generate the phrase-based translation models.

This system has been applied to the translation of WordNet (WN) [Fellbaum, 1998] glosses from English into Spanish. The experimental setting is deployed in section 2. Results are reported in section 3.

Moreover, novel ways to incorporate the information contained in the *MCR* [Atserias *et al.*, 2004] into the SMT system are being studied. Preliminary results are presented in section 4.

Finally, problems of our approach, and some possible solutions to these problems are commented in section 5.

2 Experimental Setting

As depicted in section 1, in order to build a SMT system we only need to build a *language model*, and a *translation model*, in a format that's convenient for the *Pharaoh decoder*.

Because we are translating from English into Spanish, we need a Spanish text to build the language model and an English-Spanish parallel corpus to build the translation model.

The language model must be extracted from a text as similar as possible to the target text. And because we are translating definitions, the language model has been built out from a Spanish electronic dictionary, consisting of about 300,000 definitions summing up a total of 3.5 million tokens. A trigram language model has been constructed. Linear interpolation has been applied for smoothing. Apart from that, default parameters were used.

However, for the translation model we didn't have any large enough English-Spanish parallel collection of definitions. Therefore We decided to use the European Parliament Proceedings [Koehn, 2003a] [3].

1,039,284 parallel segments ara available. That results in 28,377,767 tokens for English and 29,647,971 tokens for Spanish. A set of 327,368 segments of length between five and twenty were selected for the training of English-Spanish translation models. We used the GIZA++ default configuration (5 iterations for model 1, 4 iterations for model 3, 3 iterations for model 4, and 5 iterations for HMM model) to estimate the word-alignment

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

probabilities, i.e. $P(f|e)$ and $P(e|f)$. Furthermore, we used the Viterbi alignments (generated by GIZA++ as an intermediate result) to build the phrase-alignment probability table by means of the phrase-extract algorithm as depicted in [Och, 2002]. This algorithm takes as input a word alignment matrix and outputs a set of phrase[4] 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 language phrase are only aligned to words of the target language phrase, and viceversa.

To keep the models in a convenient size, only phrases of length between one and five have been considered. Also, phrase pairs in which the source/target phrase was more than three times longer than the target/source phrase have been ignored. Finally, phrase pairs appearing only once have been discarded, too.

Thus, we have built the language model and the translation model. Plus, we have the decoder. Apart from that, all we need is a parallel set of definitions for evaluation purposes. Fortunately, in the Spanish WordNet (version 1.6) there are 3880 glosses available. By means of the MCR we can match these to the English counterpart glosses. We'll refer to this set as *'test_set_A'*. However, these glosses are not exactly parallel but 'quasi-parallel', in the sense that they're not translations of each other although they are of course referring to the same concept.

3 Results

Several evaluation metrics have been computed, namely the General Text Matching F-measure (GTM) [Melamed *et al.*, 2003], the BLEU score [Papineni *et al.*, 2001], and the NIST score [Lin and Hovy, 2002] which have proved to correlate well with both human adequacy and fluency. All these metrics 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, we have just one reference translation.

All three measures reward longer matches. BLEU and NIST compute a word penalty factor that punishes candidate translations that are too long/short. GTM F-measure takes this into account by its own definition.

Also, BLEU and NIST don't have a clear interpretation whereas 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.

Furthermore, in our opinion, the main drawback of BLEU score is that although it works well at the document level it is not adequate at the segment level because it rewards longer matches by computing the geometric average of N-gram counts thus returning 0 when there's an n for which the n-gram match count is 0. NIST and GTM don't show this

[4] The term 'phrase' used hereafter is a little abusive since these 'phrases' are not necessarilly syntactically motivated *("a word or group of words forming a syntactic constituent with a single grammatical function")*. 'Phrase' is only referring to a sequence of words in a text. Anyhow, this term will be used for the sake of coherence with related works.

problem. NIST performs an arithmetic average instead. GTM works at the segment level by definition, rewarding longer runs by means of the *e parameter*.

In Table 1 you can see results for several translation models. The *WB-e2f* model is word-based and utilizes directly the Spanish-to-English word-alignment tables output by GIZA++. The rest of the models are phrase-based. The *PB-f2e* model uses all phrase pairs extracted as described in section 2 from the English-to-Spanish Viterbi word alignment. The *PB-*$\cup$ model considers phrase pairs belonging to the union of the English-to-Spanish and Spanish-to-English Viterbi word alignments. The *PB-Hk* model explores the space between the intersection and the union of the two Viterbi word alignments based on the *'hdiag-and'* heuristic detailed in [Koehn, 2003b]. Similarly, the *PB-Ho* model explores this same space based on the heuristic described in [Och and Ney, 2004].

	GTM-1	GTM-2	GTM-3	BLEU-1	BLEU-2	BLEU-3	BLEU-4	NIST-5
WB-e2f	**0.2622**	**0.1876**	**0.1727**	0.2387	0.1259	0.0706	0.0408	2.1132
PB-f2e	0.2551	0.1865	0.1726	0.2405	**0.1289**	**0.0730**	0.0421	**2.2202**
PB-$\cup$	0.2498	0.1843	0.1711	0.2331	0.1264	0.0712	0.0410	2.1519
PB-Hk	0.2519	0.1831	0.1692	0.2393	0.1275	0.0718	0.0419	2.1952
PB-Ho	0.2534	0.1846	0.1707	**0.2408**	0.1288	0.0728	**0.0428**	2.2192

Table 1: MT evaluation metrics on test_set_A (3880 'quasi-parallel' glosses). GTM-1, GTM-2 and GTM-3 show the GTM metric fore different values of e (e=1.0, e=2.0 and e=3.0, respectively). BLEU-1, BLEU-2, BLEU-3 and BLEU-4 show the accumulated BLEU score for different n-gram levels. Finally, NIST-5 shows the accumulated NIST score for 5-grams.

No translation model is consistently better than others when having into account all metrics. According to GTM, the word-based model is always best. However, according to BLEU and NIST scores, phrase-based models are better. Although there's no clear winner among them, according to BLEU-4, it seems that using the heuristic in [Och and Ney, 2004] works best. Regarding the NIST score, it seems that no heuristic is necessary to achieve the best result, although differences are minor.

Anyway, results, at a first glance, seem quite low compared to the performace of the same system on a set of 8490 unseen sentences from the European Parliament Proceedings. See Table 2.

To understand better these results an error analysis has been performed based on the GTM measure. This metric allows us to analyze results at the segment level, according to their F-measure. We've used the GTM-2 F-measure (e = 2) to inspect the translations output by the system using model *WB-e2f*. Table 3 reflects the distribution of the segments in test_set_A, according to the F-measure achieved.

A great number of translations scored 0 (17.76%), possibly indicating disastrous translation, whereas only 5 translations scored 1 (0.1%), sure indicator for perfect translation (See Table 4). Moreover, most translations (96.8%) achieve an F-measure lower than 0.5. And 80.21% of the translations score lower than 0.3. But, what does this mean? Are these

	GTM-1	GTM-2	GTM-3	BLEU-1	BLEU-2	BLEU-3	BLEU-4	NIST-5
EU-hk	0.5885	0.3567	0.3193	0.5963	0.4426	0.3439	0.2725	7.2477

Table 2: MT evaluation metrics for the phrase-base SMT system, on a set of 8490 unseen sentences belonging to the European Parliament Proceedings. GTM-1, GTM-2 and GTM-3 show the GTM metric fore different values of e (e=1.0, e=2.0 and e=3.0, respectively). BLEU-1, BLEU-2, BLEU-3 and BLEU-4 show the accumulated BLEU score for different n-gram levels. Finally, NIST-5 shows the accumulated NIST score for 5-grams.

translations actually so bad?

In our opinion, the answer is not, not at all. By carefully inspecting the 689 cases of totally calamitous translations (GTM-2 F-measure of 0.0) that have been observed (table 3) we found at least three reasonable arguments that will help the reader to surely look at these results with indulgence. Some translation examples may be seen in Table 5.

In first place, translation models are built on texts diametrically different from the ones under study (parliament proceedings vs. dictionary definitions). Glosses have different lexica. For instance, in case 115 the translation model did not have any translation for *'soak'* or *'bathtub'*. Besides, glosses have remarkably different syntax. In most cases, glosses are not exactly well-formed in the sense that glosses are not sentences but clauses (e.g *'transported by water'*) or even phrases (e.g *'a human being'*).

Second, the language model is built on a Spanish dictionary which definition sublanguage is also different from WordNet's. There are definitions that systematically begin in the same manner (e.g. *'the act of ...'*) that are systematically translated different than expected. See cases 350 and 485 in Table 5.

Third, as advanced in section 2, glosses are not always parallel, but *'quasi-parallel'*. Although conveying the same meaning, in a number of cases, translations are rather free with respect to the sentence structure and length. Observe in Table 3 how translation output references are, in average, 20% shorter than the source. This problem is clear in cases like 10, 464, 475, 591, 797, 1144, 1337, or 3865 (table 5). In cases 10, 475, 591, 797, 1144 and 1337 the reference is shorter than the input. It may either happen that part of the meaning in the source is not in the reference (e.g. 1144, 1337), or, that the reference somehow paraphrases the source (e.g. 10, 475, 591, 797). In cases 464 and 3865 the reference is longer because the definition includes an example.

Although all these cases scored an F-measure of 0, for many of them one can manage to understand most of the meaning they convey. Indeed, 475 and 3865 are cases of *'perfect'* translations that scored 0. Some other cases (10, 797, 1337, 350) show only minor grammatical deficiencies.

In other cases, like 1658, 1723, 2634, the system simply is unable to produce any coherent output. Usually what happens is that the lexical choice is incorrect, like in case 1658 where *'love'*, which is a noun, is translated into *'encantaría'*, which is a right translation for *'love'* as a verb but not as a noun. *'amor'* should have been chosen instead.

F-measure	#examples	proportion	$S_{avglength}$	$T_{avglength}$	$R_{avglength}$
0	689	0.1776	6.7881	6.7881	4.9390
1	5	0.0013	4.4000	4.4000	4.4000
$0.0 - 0.5$	3756	0.9680	9.3392	9.3392	7.5469
$0.0 - 1.0$	3880	1	9.2397	9.2397	7.4740
$0.0 - 0.1$	893	0.2302	8.9832	8.9832	6.4502
$0.1 - 0.2$	1311	0.3379	10.7895	10.7895	8.8978
$0.2 - 0.3$	908	0.2340	9.0914	9.0914	7.4725
$0.3 - 0.4$	458	0.1180	7.4738	7.4738	6.6463
$0.4 - 0.5$	186	0.0479	6.6290	6.6290	5.8710
$0.5 - 0.6$	68	0.0222	6.5588	6.5588	5.5441
$0.6 - 0.7$	28	0.0072	5.9643	5.9643	4.9286
$0.7 - 0.8$	17	0.0044	6.1765	6.1765	5.1176
$0.8 - 0.9$	6	0.0015	5.3333	5.3333	4.8333
$0.9 - 1.0$	5	0.0013	4.4000	4.4000	4.4000

Table 3: Distribution of examples in 'test_set_A' according to their F-measure. *F-measure* determines the exploration interval. *#examples* and *proportion* reflect the number of examples in the interval and the proportion compared to the whole set of examples, respectively. $S_{avglength}$, $T_{avglength}$ and $R_{avglength}$ show the average length of the source, target and reference, respectively, for the examples in the given interval.

This happens because the system only considers information related to the lexical units and does not utilize any further linguistic information (part-of-speech, syntactic constituents, word senses, semantic roles, etc).

4 Using WordNet

As a first manner to provide the system with outer knowledge, we explored the possibilty offered by *Pharaoh* to annotate the input with alternative translation options via XML-markup:

```
a <NN english="hombre|humano|ser humano">human being</NN>
```

This is telling the system that *'human being'* can be translated as *'hombre'*, *'humano'* and *'ser humano'*. It's therefore a very natural and convenient way to feed the system with information from outer sources of knowledge, such as WordNet. Additionally, translation probabilities for every translation option may be provided.

Using the *MCR* it is easy to retrieve all the Spanish lexicalizations of a given English word. For every synset in the English WordNet in which a given English word participates, all the variants of the synset counterpart in the Spanish WordNet are selected.

#gloss	Source	Target	Reference
1624	the study of the physical properties of light	estudio de las propiedades físicas de la luz	estudio de las propiedades físicas de la luz
1884	formal accusation of a crime	acusación formal de un delito	acusación formal de un delito
3237	common ownership	propiedad común	propiedad común
3434	lack of independence or self-sufficiency	falta de independencia o autosuficiencia	falta de independencia o autosuficiencia
3720	100 years	cien años	cien años

Table 4: Examples of 'perfect' translations (achieving a GTM-2 F-measure of 1.0). '#gloss' identifies the gloss in the test_set. 'Source' and 'Target' refer to the input and the output of the system, respetively. 'Reference' shows to the expected output.

Because this number of lexicalizations may be quite big (tens of variants), we considered using the eXtended WordNet[5] (XWN). This resource contains WordNet 2.0 English glosses as well as additional linguistic information (PoS-tagging, Parsing and Word Sense, etc.). So far, we've only used the sense information. This allows to reduce the number of lexicalizations by looking up only the variants of the specific synset. We did so only for *'gold'* quality tokens.

We built test_set_B by matching XWN glosses against the 3880 glosses in test_set_A. 3778 glosses matched. Remember that test_set_A is based on WordNet version 1.6 while test_set_B is based on WordNet version 2.0.

In Table 6 you can see translation results using four different types of information at the input. The language model is the same as in section 3. The translation model is in all cases the word-based model built from the Spanish-to-English word-alignment tables output by GIZA++.

The WB_0 experiment is our baseline. It takes as input the regular test_set_B without enrichment. No alternative lexicalizations are provided. Hence, the SMT system uses only its own information (the translation model learned from the European Parliament). See a piece of input in Table 7.

The *WB-wn* experiment takes as input test_set_B annotated with lexicalizations. All the alternative translations are assigned the same probability. See Table 8.

The *WB-wn1* experiment takes as input test_set_B annotated with lexicalizations, and translation scores. In fact, these scores are counts of the number of senses of the given source word that lexicalized as the given target word. See Table 9.

The *WB-wn2* experiment takes as input the same information as in *WB-wn1* this time normalizing the scores so they behave as probabilities. See Table 10.

[5]eXtended WordNet is freely available for public use at http://xwn.hlt.utdallas.edu/index.html

No improvement is reported. However, additional translation probabilities heuristically estimated from WordNet may help. The main reason for the decrease in performance is that only lemmas were provided as translation options. See some examples in Table 11 (conflicting words are highlighted). Spanish morphology is much richer than English's. So, in most cases the chance of finding a correct translation was null. Besides, we trusted word sense disambiguation as supplied by XWN, which is far from an acceptable level of performance, thus introducing a considerable amount of noise.

For instance, in case 972, when translating *'1st'* into *'primero'*, the gender is wrong (*primero/primera*) because in Spanish the gender of an adjective must match the gender of the noun it modifies. In this case *'primero'* is masculine whereas *'vértebra'* is femenine. The cause of this translation error is that *'primero'* is the lemma of *'primera'*. All word forms should have been considered instead of the lemma alone.

In case 1069, when translating *'studies'* into *'estudiar'*, the verbal form is wrong (*estudiar/estudia)*. Again, the reason is that *'estudiar'* is the lemma of *'estudia'*. Case 1069 is also an example of bad lexical choice. The word *'plants'* is wrongly translated into *'factoría'* (*'factory'*), which is a good translation for plant but not in this case (*'planta'* would be the right choice). See translation options for *'plants'* in Table 12. Also the gender was wrong, because *'plants'* is plural whereas *'factoría'* is singular.

However, Table 13 suggests that WordNet may be useful in many cases (see highlighted words). For instance, cases 1469 and 2146 show enhancement in lexical choice adequacy. 116, 780, 1826, and 3378 are cases in which a bad lexical choice is corrected using information from the MCR. Specially paradigmatic is the case 1826 in which the word *'bank'*, previously wrongly translated as *'Banco'* (economics), is now correctly translated as *'orilla'*. In cases 2911, 2986, 2691, 3170, and 3315 there were some words unknown to the translation model, which have been retrieved from the MCR.

5 Discussion and Further Work

MT quality achieved by the system on the translation of WordNet glosses is still low. However, as commented in section 3, evaluation has not been exactly fair. In order to perform a fair evaluation, we must redefine the test_set. We could, for instance, discard glosses for which the length of the reference translation is too different from the source. That would fight against an excessive amount of 'quasi-parallelism' at the cost of a decrease in statistical significance caused by the decrement in the size of the test set.

However, MT Evaluation is still a problem under discussion and continuous development. We consider to incorporate new metrics, such as ROUGE [Lin and Och, 2004]. Moreover, we are also considering to evaluate the system on human judgements so as to measure its actual usability.

As to improving of the system, we are studying the use of additional linguistic information other than lexical units. We plan to utilize part-of-speech tagging, chunking, named entity recognition, semantic role labeling, and clause splitting.

So far, we have used the MCR to provide the system with outer knowledge because,

from a conceptual point of view, the idea of enriching WordNet using WordNet itself results very attractive. However, no gain in MT quality has been observed. This may be due to the fact that translation options provided were not word forms but lemmas. We suspect that using word forms instead of lemmas could highly increase the system performance. Of course, this would generate many more translation options. Therefore, we think it may be useful to consider this information only in certain cases:

- unknown[6] words
- words for which we are confident to know the precise meaning (e.g *'gold'* quality).
- words up to a certain degree of polysemy
- words belonging to a certain PoS (noun/verb/adjective/adverb)

Of course, a natural evolution of this approach is to use a Word Sense Disambiguation system to enrich the input instead. Remember we have simulated this by utilizing eXtended WordNet.

Another path we are exploring is the tuning of the *Pharaoh* parameters that adjust the importance the different probabilities that govern the search. Recall the problem of the difference in length discussed in section 3. This suggests that tuning the Pharaoh *word penalty* parameter may lead to better results.

In general, there are 4 important parameters to adjust:

- λ_{lm} language model probability
- λ_{ϕ} translation model probability
- λ_{d} distortion probability
- λ_{w} word penalty factor

We believe our system can serve as a starting point to aid human translation of WordNet glosses. Hence, WordNet versions other than English could be filled up with glosses with little effort.

[6]words that have not been seen in the training of the translation model.

#gloss	Source	Target	Reference
10	a living organism lacking the power of locomotion	falta de un organismo vivo el poder locomotion	vida vegetal
115	you soak your body in a bathtub	que soak su cuerpo en un bathtub	acción y efecto de bañarse
350	the act of teaching at school	acción de enseñar a la escuela	educación escolar
464	strain oneself more than is healthy	que tiene buena disposición sí más	esforzarse con afán para conseguir alguna cosa : se mata a leer .
475	eat a large amount of food quickly	comer una gran cantidad de comida rápidamente	engullirse ávidamente
485	the act of showing regard for others	acción de pruebas respecto de los demás	buenos modales en el comportamiento en grupo
591	related by blood ; not adopted	relación de parentesco ; no aprueba	" Hijo natural "
797	clearly superior or having the attributes of a winner	superiores , o con la atributos de una vencedor	que destaca claramente en algo
1144	reduces frequency distortion	variedad frecuencia distorsión	(música) (tecnología)
1337	equipment used to broadcast radio or tv signals	aparato utilizado para transmitir señales radio o tv	(tecnología)
1658	a story dealing with love	que una cosa con encantaría	historia de amor
1723	an official report (usually sent in haste)	oficial general envió mismo precipitación en un (.	comunicación hecha por vía rápida
2634	a church official	sufrir una iglesia	oficial eclesiástico
3865	a newspaper	un periódico	publicación periódica monotemática (cultura , política ...)

Table 5: Examples of translations achieving a GTM-2 F-measure of 0.0. '#gloss' identifies the gloss in the test_set. 'Source' and 'Target' refer to the input and the output of the system, respetively. 'Reference' shows the expected output.

	GTM-1	GTM-2	GTM-3	BLEU-1	BLEU-2	BLEU-3	BLEU-4	NIST-5
WB_0	0.2627	0.1877	0.1728	0.2388	0.1259	0.0706	0.0408	2.1145
WB-wn	0.2342	0.1704	0.1575	0.2104	0.1064	0.0572	0.0319	1.8183
WB-wn1	0.2431	0.1753	0.1616	0.2183	0.1116	0.0606	0.0343	1.9267
WB-wn2	0.2432	0.1754	0.1617	0.2184	0.1117	0.0606	0.0343	1.9283

Table 6: MT evaluation metrics on test_set_B (3778 'quasi-parallel' WordNet (enriched) glosses). GTM-1, GTM-2 and GTM-3 show the GTM metric fore different values of e (e=1.0, e=2.0 and e=3.0, respectively). BLEU-1, BLEU-2, BLEU-3 and BLEU-4 show the accumulated BLEU score for different n-gram levels. Finally, NIST-5 shows the accumulated NIST score for 5-grams.

```
...
a human being
...
a point or extent in space
...
```

Table 7: A piece of input without using WordNet.

```
...
a <NN english="hombre|humano|ser humano">human being</NN>
...
a <NN english="punto">point</NN> or <NN english="alcance|extensión">
extent</NN> in <NN english="blanco|espacio|espacio en blanco|espacio
exterior">space</NN>
...
```

Table 8: A piece of input using WordNet. The 'english' XML attribute containing the list of translation candidates receives this name only for historical reasons.

```
...
a <NN english="hombre|humano|ser humano" prob="1|1|1">human being</NN>
...
a <NN english="punto" prob="1">point</NN> or
<NN english="alcance|extensión" prob="1|2">extent</NN>
in <NN english="blanco|espacio|espacio en blanco|espacio
exterior" prob="1|4|1|1">space</NN>
...
```

Table 9: A piece of input using WordNet, plus scores extracted from WordNet according to the number of senses of the given word that lexicalized in the same manner.

```
...
a <NN english="hombre|humano|ser humano" prob="0.3333|0.3333|0.3333">
human being</NN>
...
a <NN english="punto" prob="1">point</NN> or <NN english="alcance
| extensión" prob="0.3333|0.6666">extent</NN> in
<NN english="blanco|espacio|espacio en blanco|espacio exterior"
prob="0.1429|0.5714|0.1429|0.1429">space</NN>
...
```

Table 10: A piece of input using WordNet, plus same scores now transformed into probabilities.

#gloss	Source	Target	Reference
423	**sends** or **receives** radio or television signals	aparato eléctrico que radio o televisión **transmitir** o **recibir** una señal	aparato que **envía** o **recibe** señales de radio o televisión
972	the **1st** cervical vertebra	el **primero** vértebra cervical	**primera** vértebra cervical
1069	the branch of biology that **studies plants**	rama de la biología que **estudiar factoría**	rama de la biología que **estudia** las **plantas**
2576	someone who **drinks** liquor repeatedly in small quantities	persona que **beber** licor continuamente en **poco** cuantía	persona que **bebe** en pequeñas cantidades
3223	one of the **twelve** divisions of **the** calendar year	una de las doce **parte** de **la** año civil	cada una de las doce **partes** en que se divide **el** año

Table 11: Examples of translations using WordNet to enrich the input. Conflicting words are highlighted. '#gloss' identifies the gloss in the test_set. 'Source' and 'Target' refer to the input and the output of the system, respetively. 'Reference' shows the expected output.

```
the <NN english="afluente|brazo|rama|subdivisión" prob="0.25
|0.25|0.25|0.25">branch</NN> of <NN english="biología" prob="1">
biology</NN> that <VBZ english="aprender|considerar|contemplar|
cursar|estimar|estudiar|estudiarse|juzgar|meditar" prob="0.0909
|0.0909|0.0909|0.0909|0.0909|0.2727|0.0909|0.0909|0.0909">studies
</VBZ> <NNS english="factoría|flora|fábrica|planta" prob="0.2
|0.2|0.2|0.4">plants</NNS>
```

Table 12: WN-aided translation input for case 1069.

#gloss	Source	Target	Target-WN	Reference
ENHANCING LEXICAL CHOICE				
1469	**piece** of **pork**	**pieza** de **cerdo**	**pedazo** de **carne de cerdo**	**trozo** de **carne de cerdo**
2146	the **wife** of a **crown** prince	**mujer** de un príncipe la **corona**	el **esposa** de un príncipe **heredero**	**esposa** de un príncipe
CORRECTING BAD LEXICAL CHOICE				
116	flying in a **balloon**	volar en un **hinchado**	volar en un **globo**	acción de volar en un **globo aerostático**
780	a higher **deck**	**vuelo** un superior	**cubierta** superior NULL	**cubierta** superior
1826	the **bank** of a river	el **Banco** de un río	la **orilla** de un río	**orilla** de un río
3378	holding first place in a **contest**	en primer lugar en una **belleza**	hacer primer lugar en una **competición**	que ha obtenido el primer lugar en una **competición**
UNKNOWN WORDS				
2911	absence of light or **illumination**	falta de luz o **illumination**	ausencia de luz o **iluminación**	ausencia de luz
2986	**inflammation** of the **conjunctiva** of the eye	**curarnos** de la **conjunctiva** de los ojos	**inflamación** de la **conjuntiva** de la vista	**inflamación** de la **conjuntiva**
2691	any part of a plant or **fungus**	cualquier parte de una planta o **fungus**	cualquier parte de una planta o **hongo**	parte de una planta o de un **hongo**
3170	a **span** of 1000 years	una de mil años **2005**	**espacio** de mil años NULL	**espacio** temporal de mil años
3315	cover with **gravel**	cubrir con gravel	cubrir con **gravilla**	cubrir con**grava**

Table 13: Examples of translations using WordNet to enrich the input. Conflicting words are highlighted. '#gloss' identifies the gloss in the test_set. 'Source', 'Target' and 'Target-WN' refer to the input, output without using WN and output using WN, respetively. 'Reference' shows the expected output.

References

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

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

[Brown *et al.*, 1990] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, , and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85, 1990.

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

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

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

[Koehn, 2003b] Philipp Koehn. *Noun Phrase Translation*. PhD thesis, University of Southern California, 2003.

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

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

[Lin and Och, 2004] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statics. In *Proceedings of ACL'04*, 2004.

[Melamed *et al.*, 2003] I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and recall of machine translation. In *Proceedings of HLT/NAACL'03*, 2003.

[Och and Ney, 2003] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[Och and Ney, 2004] Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

[Och, 2002] Franz Josef Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen, Germany, 2002.

[Papineni *et al.*, 2001] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation, ibm research report, rc22176. Technical report, IBM T.J. Watson Research Center, 2001.

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

[Stolcke, 2002] Andreas Stolcke. Srilm - an extensible language modeling toolkit. In *Proceedings of ICSLP'02*, 2002.

[Weaver, 1955] Warren Weaver. *Translation (1949)*. Machine Translation of Languages. MIT Press, Cambridge, MA, 1955.