Using POS Information for Statistical Machine Translation into Morphologically Rich Languages

Nicola Ueffing and **Hermann Ney**
Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen – University of Technology
{ueffing,ney}@cs.rwth-aachen.de

Abstract

When translating from languages with hardly any inflectional morphology like English into morphologically rich languages, the English word forms often do not contain enough information for producing the correct fullform in the target language. We investigate methods for improving the quality of such translations by making use of part-of-speech information and maximum entropy modeling. Results for translations from English into Spanish and Catalan are presented on the LC-STAR corpus which consists of spontaneously spoken dialogues in the domain of appointment scheduling and travel planning.

1 Introduction

In this paper, we address the question of how part-of-speech (POS) information can help improving the quality of Statistical Machine Translation (SMT). One of the main problems when translating from a language with hardly any inflectional morphology (which is English in our experiments) into one with richer morphology (here: Spanish and Catalan) is the production of the correct inflected form in the target language. We introduce transformations to the English string that are based on the part-of-speech information and show how this knowledge source can help SMT. Systematic evaluations will show that the quality of the generated translations is improved.
The transformations we apply are the following:

Treatment of verbs In Catalan and Spanish, the pronoun before a verb is often omitted and instead, the person is expressed via the ending of the verb. The same holds for future tense and for the modes expressed through 'would' and 'should' in English. Since this makes it hard to generate the correct translation of a given English verb, we propose a method resulting in English word forms containing sufficient information.

Question inversion In English, interrogative phrases have a word order that is different from declarative sentences: Either an auxiliary 'do' is inserted or the order of verb and pronoun is inverted. Since this is different in Spanish and Catalan, we modify the word order in English to make it more similar to the Spanish/Catalan one and to help the verb treatment mentioned above.

The paper is organized as follows: Related work is treated in Section 2. In Section 3, we shortly review the statistical approach to machine translation. Then, we introduce the transformations that we apply to the less inflected language of the two under consideration (namely English) in Section 4. After describing the maximum entropy approach and the training procedure we use for the statistical lexicon in Section 5, we present results on the trilingual LC-STAR corpus in Section 6. Then, we conclude and present ideas about future work in Section 7.

2 Related Work

Publications dealing with the integration of linguistic information into the process of *statistical* machine translation are rather few although this had already been suggested in (Brown et al., 1992). (Nießen and Ney, 2001b) introduce hierarchical lexicon models including baseform and POS information for translation from German into English. Information contained in the German entries that are not relevant for the generation of the English translation are omitted. Unlike this, we investigate methods for enriching English with knowledge to help selecting the correct fullform in a morphologically richer language.

(Nießen and Ney, 2001a) propose reordering operations for the language pair German–English that help SMT by harmonizing word order between source and target. The question inversion we apply was inspired by this; nevertheless, we do not perform a full morpho-syntactic analysis, but make use only of POS information which can be obtained from freely available tools.

(Garcia-Varea et al., 2001) apply a maximum entropy approach for training the statistical lexicon, but do not take any linguistic information into account.

The use of POS information for improving statistical alignment quality is described in (Toutanova et al., 2002), but no translation results are presented.

3 Statistical Machine Translation

The goal of machine translation is the translation of an input string $s_1, \ldots, s_J$ in the source language into a target language string $t_1, \ldots, t_I$. We choose the string that has maximal probability given the source string, $Pr(t_1^I|s_1^J)$. Applying Bayes' decision rule yields the following criterion:

$$
\begin{aligned}
& \arg\max_{t_1^I} Pr(t_1^I|s_1^J) \\
= \quad & \arg\max_{t_1^I} \{Pr(t_1^I) \cdot Pr(s_1^J|t_1^I)\} \qquad (1)
\end{aligned}
$$

Through this decomposition of the probability, we obtain two knowledge sources: the translation and the language model. Those two can be modelled independently of each other.

The correspondence between the words in the source and the target string is described by alignments that assign target word positions to each source word position. The probability of a certain target language word to occur in the target string is assumed to depend basically only on the source words aligned to it.

The search is denoted by the $\arg\max$ operation in Eq. 1, i.e. it explores the space of all possible target language strings and all possible alignments between the source and the target language string to find the one with maximal probability.

The input string can be preprocessed before being passed to the search algorithm. If necessary, the inverse of these transformations will be applied to the generated output string. In the work presented here, we restrict ourselves to transforming only one language of the two: the source, which has the less inflected morphology.

For descriptions of SMT systems see for example (Germann et al., 2001; Och et al., 1999; Tillmann and Ney, 2002; Vogel et al., 2000; Wang and Waibel, 1997).

4 Transformations in the Less Inflected Language

When translating from English into languages with a highly inflected morphology, the production of the correct fullform often causes problems. Our experience on several corpora shows that the error rate of a translation from English into morphologically richer languages decreases by 10% relative if we aim at producing only the correct baseform instead of the fully inflected word. The transfer of the meaning expressed in the baseform is easier than deciding on the correct inflected form.

4.1 Treatment of Verbs

Especially the translation of verbs is difficult since there are many different inflections in Spanish and Catalan whereas there are only few in English. Moreover, the pronouns and modals are often omitted in Spanish and Catalan and this information is expressed through the suffix. This makes it very hard for word-based systems to generate the correct inflection from the English verb which does not contain sufficient information. Thus, several English words will have to be aligned to the Spanish or Catalan verbs. This process is relatively difficult for the algorithm and causes noise in the statistical lexicon if English pronouns are regarded as translations of Spanish or Catalan verbs. In order to enrich the English verb with the needed information, we combine pronouns and/or modals with following verbs and treat those combinations as 'new' fullform words in English. Thus we can obtain the information needed to select the correct verb form in the target language from one single English word. The identification of English pronouns, modals and verbs was done by POS tagging applied to the English part of the corpus.

We decided to transform the source language instead of the target language, because in this case we need only the POS tags of the source language as additional knowledge source and nothing else. Another possible approach would have been to split the suffix in the target language (e.g. 'está' into 'estar P3S'). This would require postprocessing tools that are able to generate the correct verb form from the baseform and the person and tense information.

Table 1 gives examples of words that have been spliced to form new entries of the English lexicon. For example, we splice the phrase 'you think' to form the single entry 'you_think' which contains sufficient information for producing the correct Spanish verb form 'crees' or the Catalan 'creus'. Similarly, the modal auxiliaries can be added as well, like in the entry 'you_will_have' which is much better suited for being translated into 'tendrás' (Spanish) or 'tindràs' (Catalan) than the verb 'have' alone. Moreover, in a single word based lexicon, three single entries would have to be added for the translation of 'you will have' into 'tendrás': (you,tendrás), (will,tendrás) and (have,tendrás), which spreads the translation probability over far too many entries and makes the probability distribution unfocused.

As the last example in Table 1 shows, 'you can go' is spliced only into two words instead of one in order to better match the Spanish/Catalan form.

4.2 Question Treatment

In English interrogative phrases, either an auxiliary 'do' is inserted or the order of verb and pronoun is inverted. The auxiliary 'do' does not carry information that is relevant when translating into Spanish or Catalan. Thus, we can remove it from the sentence without harming the translation process (as described in (Nießen and Ney, 2001a) for the language pair German–English). However, we do not remove a question supporting 'do' in past tense, i. e. 'did' is kept in the phrase, because this is the only word containing the tense information. Afterwards, we can merge the pronoun and verb as depicted in Table 2: 'did you go' is transformed into 'you_did go'. We do not splice 'you_did' and 'go', because the English simple past is translated into present perfect in Catalan; and it is very likely to be translated into present perfect in Spanish, especially in colloquial language as it is present in this task. The form 'you_did go' is well suited to be translated into the Spanish 'has ido' or the Catalan 'has anat'.

Table 1: Examples of spliced words in the English vocabulary

original	POS tags	spliced words
you go	PRP VBP	you_go
you went	PRP VBD	you_went
you think	PRP VBP	you_think
you will have	PRP MD VB	you_will_have
you can go	PRP MD VB	you_can go

If there is no question supporting 'do' and the order of pronoun and verb is inverted – see the example 'how are you?' in Table 3 – we first swap the two words and then perform the splicing step. This is done in order to avoid having two lexical entries with the same translation: for example, 'you_are' and the interrogative 'are_you' both have the same translation in Spanish or Catalan, respectively.

Table 3 presents examples of transformed English questions. Comparing them to the Spanish and Catalan reference, we see that it is easier to find a word-to-word mapping for the modified English sentences.

5 Maximum Entropy Training

If we merge the pronouns/modals and verbs as described above, it might happen that the verb itself (or one of its inflections) has never been seen in training except from its appearance in the new entries in the lexicon which result from the splic-

Table 2: Examples of spliced words in the English vocabulary after question inversion

original	POS tags	spliced words
do you go	VBP PRP VB	you_go
did you go	VBD PRP VB	you_did go
have you gone	VBP PRP VBN	you_have gone
will you go	MD PRP VB	you_will_go
can you go	PRP MD VB	you_can go

Table 3: Examples of transformed English sentences

Original	how are you ?
Question Inversion	how you are ?
Verb Treatment	how you_are ?
Catalan Sentence	com está ?
Spanish Sentence	¿ cómo estás ?
Original	or do you think we want to stay [. . .] ?
Question Inversion	or you think we want to stay [. . .] ?
Verb Treatment	or you_think we_want to stay [. . .] ?
Catalan Sentence	o creu que voldrem quedar-nos [. . .] ?
Spanish Sentence	¿ o cree que querremos quedarnos [. . .] ?
Original	did you say the eighteenth ?
Question Inversion	you did say the eighteenth ?
Verb Treatment	you_did say the eighteenth ?
Catalan Sentence	has dit el divuit ?
Spanish Sentence	¿ has dicho el dieciocho ?

ing operation. This makes it impossible to translate the verb itself, because it is then unknown to the system. The same holds for combinations of pronouns and verbs that are unseen in training, e. g. the training corpus contains the bigram 'I went', but not the one 'she went'. In order to overcome this problem, we train our lexicon model using maximum entropy.

5.1 The Maximum Entropy Approach

The maximum entropy approach (Berger et al., 1996) presents a powerful framework for the combination of several knowledge sources. This principle recommends to choose the distribution which preserves as much uncertainty as possible in terms of maximizing the entropy. The distribution is required to satisfy constraints, which represent facts known from the data. These constraints are expressed on the basis of feature functions $h_m(s,t)$, where (s,t) is a pair of source and target word. The lexicon probability of a source word given the target word has the following functional form

$$p(s|t) = \frac{1}{Z(t)} \exp\left[\sum_m \lambda_m h_m(s,t)\right]$$

with the normalization factor

$$Z(t) = \sum_{s'} \exp\left[\sum_m \lambda_m h_m(s',t)\right] \quad ,$$

where $\Lambda = \{\lambda_m\}$ is the set of model parameters with one weight λ_m for each feature function h_m. The features we use in our model are

- a lexical feature (for the entries of the transformed vocabulary):

$$h_{s',t'}(s,t) = \delta(s,s') \cdot \delta(t,t')$$

- the verb contained in a transformed lexicon entry (e.g. 'go' for 'you_go' or 'you_will_go):

$$h_{s',v}(s,t) = \delta(s,s') \cdot Verb(t,v) \quad ,$$

where

$$Verb(t,v) = \begin{cases} 1, & \text{if } t \text{ contains the verb } v \\ 0, & \text{otherwise} \end{cases}$$

This enables us to translate the verb alone even if it occurs in the training corpus only as a spliced entry.
For an introduction to maximum entropy modeling and training procedures, the reader is referred to the corresponding literature, for instance (Berger et al., 1996) or (Ratnaparkhi, 1997).

5.2 Training

We performed the following training steps:

- transform the English (= source language) part of the corpus as described in Sections 4.1 and 4.2
- train the statistical translation system using this modified source language corpus [1]
- with the resulting alignment, train the lexicon model using maximum entropy with the features described in Section 5.1

This training can be performed using converging iterative training procedures like described by (Darroch and Ratcliff, 1972) or (Della Pietra et al., 1997) [2]. The basic training procedures for the translation system and the language model need not be changed.

5.3 Translation process

For translation, we can use an SMT system where the search algorithm does not have to be modified. Before the translation process, we transform the input in the same way as the training corpus before training the alignment (see Section 5.2). We simply have to exclude those words from splicing where the splicing operation yields an unknown word.

[1]This training was done using the GIZA++ toolkit which can be downloaded from http://www-i6.informatik.rwth-aachen.de/~och/software/GIZA++.html

[2]We made use of the toolkit YASMET which can be downloaded from http://www-i6.informatik.rwth-aachen.de/~och/software/YASMET.html

6 Results

6.1 Corpora

We performed experiments on the trilingual corpus which is successively built within the LC-STAR project. It comprises the languages English, Spanish and Catalan, whereof we used English as source and Spanish and Catalan as target languages. At the time of our experiments, we had about 13k sentences per language available; the statistics are given in Table 4.
The corpus consists of transcriptions of spontaneously spoken dialogues. Thus, the sentences often lack correct syntactic structure. The domain of this task is appointment scheduling and travel arrangements.
The POS information for the English part of the corpus was generated using the Brill tagger[3].
As Table 4 shows, the splicing operation increases the cardinality of the English vocabulary as well as the number of singletons significantly. Nevertheless, they are still below those numbers for Spanish and Catalan.

6.2 Evaluation Metrics

The quality of the output of our machine translation system is measured automatically by comparing the generated translation to a given reference translation. The two following criteria are used:

- WER (**w**ord **e**rror **r**ate):
  The word error rate is based on the Levenshtein distance. It is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated string into the reference string. Since some sentences in the develop and test set occur several times with different reference translations (which holds especially for short sentences like 'okay, good-bye'), we calculate the minimal distance to this set of references as proposed in (Nießen et al., 2000).

- BLEU (**bil**ingual **e**valuation **u**nderstudy):
  (Papineni et al., 2002) have proposed a

[3]The Brill tagger can be downloaded from http://www.research.microsoft.com/users/brill/

Table 4: Statistics of the training, develop and test set of the English-Spanish-Catalan LC-STAR corpus (*number of words without punctuation marks)

		English		Spanish	Catalan
		Original	Transformed		
Training	Sentences	13 352			
	Words	123 454	114 099	118 534	118 137
	Words*	101 738	92 383	96 997	96 503
Vocabulary	Size	2 154	2 776	3 933	3 572
	Singletons	790 (37%)	1 165 (42%)	1 844 (47%)	1 658 (47%)
Develop	Sentences	272			
	Words	2 267	2 096	2217	2211
	Unknown Words	21	22	34	34
Test	Sentences	262			
	Words	2 626	2 460	2 451	2 470
	Unknown Words	17	18	30	35

method of automatic machine translation evaluation, which they call "BLEU". It is based on the notion of modified n-gram precision, for which all candidate n-gram counts in the translation are collected and clipped against their corresponding maximum reference counts. These clipped candidate counts are summed and normalized by the total number of candidate n-grams. Since BLEU expresses quality, we determine $100-\text{BLEU}$ to transform it into an error measure.

Although these measures are only approximations, they seem to be sufficient at the present level of performance of machine translation systems.

6.3 Experimental Results

We compared the two statistical lexica obtained from the baseline system and from the maximum entropy training on the transformed corpus. For the baseline lexicon, we observed an average of 5.82 Catalan translation candidates per English word and 6.16 Spanish translation candidates. These numbers are significantly reduced in the lexicon which was trained on the transformed corpus using maximum entropy: there, we have an average of 4.20 for Catalan and 4.46 for Spanish. Especially for (nominative) English pronouns (which have many verbs as translation candidates in the baseline lexicon), the number of translation candidates was substantially scaled down by a factor around 4. This shows that our method was successful in producing a more focused lexicon probability distribution.

We performed translation experiments with an implementation of the IBM-4 translation model (Brown et al., 1993). A description of the system can be found in (Tillmann and Ney, 2002).

Table 5 presents an assessment of translation quality for both the language pairs English–Catalan and English–Spanish. We see that there is a significant decrease in error rate for the translation into Catalan. This change is consistent across both error rates, the WER and $100-\text{BLEU}$.

For translations from English into Spanish, the improvement is less substantial. A reason for this might be that the Spanish vocabulary contains more entries and the ratio between fullforms and baseforms is higher: 1.57 for Spanish versus 1.53 for Catalan[4]. This makes it more difficult for the system to choose the correct inflection when generating a Spanish sentence. We assume that the extension of our approach to other word classes than verbs will yield a quality gain for translations into Spanish.

Table 6 shows several sentences from the English LC-STAR develop and test corpus that were trans-

[4]The lemmatization of Spanish and Catalan was produced using the analyser from UPC Barcelona: MACO+ and RELAX.

Table 5: Translation error rates [%] for English–Catalan and for English–Spanish

		Develop		Test	
		WER	100-BLEU	WER	100-BLEU
Catalan	Baseline	37.6	58.2	33.0	49.2
	+ Transformations	35.0	55.1	30.8	46.6
Spanish	Baseline	35.4	57.6	32.1	48.9
	+ Transformations	35.0	55.8	31.5	47.6

lated into Catalan. We see that it is easier for the system to generate the correct verb inflection in Catalan if the verb is enriched with the pronoun. In the baseline system, it happens that words are inserted – like 'far' as translation of 'will' in the second example which is incorrect. This can be avoided by the splicing of words.

In the last example, we see that the baseline system generates one word each for the English 'I prefer' and does not find the correct translation, whereas transformations yield an accurate translation of this expression, because the spliced word contains sufficient information.

7 Conclusion and Future Work

We presented a method for improving quality of statistical machine translation from English into morphologically richer languages like Spanish and Catalan. Using POS tags as additional knowledge source, we enrich the English verbs such that they contain more information relevant for selecting the correct inflected form in the target language. The lexicon model was then trained using the maximum entropy approach, taking the verbs as additional features.

Results were given for translation from English into Spanish and Catalan on the LC-STAR corpus which consists of spontaneously spoken dialogues in the domain of appointment scheduling and travel arrangement. Our experiments show that translation quality can be significantly increased through the use of our approach: the word error rate on the Catalan development set for example decreased by 2.5% absolute.

We plan to investigate other methods of enriching the English words with information. It will be interesting to see how other word classes, e. g. nouns, can be handled in order to improve quality of translations into languages with a highly inflected morphology.

8 Acknowledgements

This work was partly supported by the LC-STAR project by the European Community (IST project ref. no. 2001-32216).

References

A.L. Berger, S.A. Della Pietra, and V.J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March.

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.D. Lafferty, and R.L. Mercer. 1992. Analysis, statistical transfer, and synthesis in machine translation. In *Proc. TMI 1992: 4th Int. Conf. on Theoretical and Methodological Issues in MT*, pages 83–100, Montréal, P.Q., Canada, June.

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

J.N. Darroch and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480.

S.A. Della Pietra, V.J. Della Pietra, and J. Lafferty. 1997. Inducing features in random fields. *IEEE Trans. on Pattern Analysis and Machine Inteligence*, 19(4):380–393, July.

I. Garcia-Varea, F.J. Och, H. Ney, and F. Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *Proc. 39th Annual Meeting of the Assoc. for Computational Linguistics - joint with EACL*, pages 204–211, Toulouse, France, July.

Table 6: Examples of English–Catalan translations with and without transformation

Source	we_exchange them and, that would be good.
Reference	*les canviem i, això estaria bé.*
Baseline	ens canviem i, això estaria bé.
Verb Treatment	les canviem i, això estaria bé.
Source	okay, and I_will, speak to you soon then.
Reference	*d' acord, i jo, parlaré amb tu aviat doncs.*
Baseline	d' acord, i jo far, parlaré amb tu aviat doncs.
Verb Treatment	d' acord, i jo, parlaré amb tu aviat doncs.
Source	I_believe, the flight is every day?
Reference	*crec, que el vol és cada dia?*
Baseline	suposo, el vol és cada dia?
Verb Treatment	crec, que el vol és cada dia?
Source	I_prefer single.
Reference	*prefereixo individual.*
Baseline	jo preferiria una individual.
Verb Treatment	prefereixo una individual.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proc. 39th Annual Meeting of the Assoc. for Computational Linguistics - joint with EACL*, pages 228–235, Toulouse, France, July.

S. Nießen and H. Ney. 2001a. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proc. MT Summit VIII*, pages 247–252, Santiago de Compostela, Galicia, Spain, September.

S. Nießen and H. Ney. 2001b. Toward hierarchical models for statistical machine translation of inflected languages. In *39th Annual Meeting of the Assoc. for Computational Linguistics - joint with EACL 2001: Proc. Workshop on Data-Driven Machine Translation*, pages 47–54, Toulouse, France, July.

S. Nießen, F.J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *Proc. of the Second Int. Conf. on Language Resources and Evaluation*, pages 39–45, Athens, Greece, May.

F.J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics*, pages 311–318, Philadelphia, PA, July.

A. Ratnaparkhi. 1997. A simple introduction to maximum entropy models for natural language processing. Technical Report 97–08, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, May.

C. Tillmann and H. Ney. 2002. Word re-ordering and DP beam search for statistical machine translation. *to appear in Computational Linguistics*.

K. Toutanova, H.T. Ilhan, and C.D. Manning. 2002. Extensions to HMM-based statistical word alignment models. In *Proc. Conf. on Empirical Methods for Natural Language Processing*, pages 87–94, Philadelphia, PA, July.

S. Vogel, F.J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H. Ney. 2000. Statistical methods for machine translation. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 377–393. Springer Verlag: Berlin, Heidelberg, New York.

Y.Y. Wang and A. Waibel. 1997. Decoding algorithm in statistical translation. In *Proc. 35th Annual Meeting of the Assoc. for Computational Linguistics*, pages 366–372, Madrid, Spain, July.