Proc. Automatic Speech Recogn. and Underst. Workshop, Cancun, Mexico, Dec. 2005.

TOWARDS INCORPORATING LANGUAGE MORPHOLOGY INTO STATISTICAL MACHINE TRANSLATION SYSTEMS

Panagiotis Karageorgakis, Alexandros Potamianos, Ioannis Klasinas

Department of Electronics and Computer Engineering
Technical University of Crete
Chania 73100, Greece
{pkarage, potam, iklasinas}@telecom.tuc.gr

ABSTRACT

In this paper, a novel algorithm for incorporating morphological knowledge into statistical machine translation (SMT) systems is proposed. First, word stems are acquired automatically for the source and target languages using an unsupervised morphological acquisition algorithm. Then a word-stem based SMT system is built and combined with a phrase-based word level SMT system using a general statistical framework. The combined lexical and morphological SMT system is implemented using late integration and lattice re-scoring. The system is then evaluated on the Europarl corpus, using automatic evaluation methods for various training corpus sizes. It is shown, that both the BLEU and NIST scores of the lexical-morphological system improve by about 14% over the baseline English to Greek translation system when using a 1M word training corpus.

1. INTRODUCTION

Statistical Machine Translation (SMT) systems has proven to be a valuable tool for the automatic translation between languages. SMT systems can be trained from untagged, parallel corpora using large amounts of bilingual (or multilingual) text documents. Evaluation of the quality of such systems may be done either by human judges, or automatically by computing metrics. Most SMT systems today employ little or no linguistic knowledge and operate purely at the the lexical level, i.e., use a “direct” translation approach. It is clear that linguistic information has the potential of improving performance of SMT systems, especially when limited amounts of parallel training data sets are available (data-sparseness problem). However, incorporating linguistic morphological, syntactic and semantic information into the statistical framework of SMT is a hard problem.

There have been recent efforts in building machine translation systems using morphological (typically word-stem) information. Germann [1] has built a statistical machine translation system between Tamil and English using a small corpus, and shows improvement in quality by the use of a simple stemmer. Driven by the same need, Sonja Niessen and Hermann Ney [2] use morphological knowledge to increase word coverage. This approach is efficient when translating from highly inflected languages like German to poor inflected languages like English. They use moprhological analysis to find the base form (lemma) of similar words, thus creating word classes which tend to be translated by the same target word.

In this paper, we propose a method of improving the translation quality of existing SMT systems, by incorporating word-stems into SMT systems. There are two major innovations in this work: First the morphological information is extracted automatically from text using a robust version of the unsupervised morphology acquisition algorithm presented in [3]. Second the stems are incorporated into the SMT system using a general statistical framework which combines a word-based and a stem-based SMT system. Both the robust morphological analyzer and lexical-morphological SMT systems are evaluated and shown to significantly improve on the state-of-the art.

The rest of this paper is organized as follows: In Section 2, the unsupervised morphological acquisition algorithm is presented. Section 3 presents the statistical framework of the combined lexical-morphological SMT system. In Section 4, the implementation of the combined SMT system is presented, using late integration and lattice combination of the word and stem SMT systems. The experimental setup and results are presented in Section 5 and 6 respectively. Finally, we conclude our work with Section 7.

2. UNSUPERVISED MORPHOLOGY ACQUISITION

Morphology is the study of the way words are built up from smaller meaning-bearing units, which are called *morphemes*. The most important morpheme of a word is the stem, or the lemma of a word, which is its root. Apart from the stem, a morpheme can be an affix (prefix, suffix, infix or circumfix), which usually provides additional meaning of some kind to the main concept that is provided by the stem.

In recent years, there has been much interest in computational models that learn aspects of the morphology of a natural language from raw or structured data. These models are of great practical interest, minimizing the expert resources or need of linguists in order to develop stemmers and analyzers. There are three distinct ways of learning a language's morphology: *supervised learning*, where the data consists of a set of pair of words, *unsupervised learning*, where the data consists of a single set of all the words in a corpus, and finally *partially supervised learning* where the data consists of two sets of words, without any indication of the relationship between the individual words. For training purposes, most SMT systems today use untagged corpora in one or more languages. Morphological analysis performed on this corpora could derive information about the morphology of the source and target languages, that could be later incorporated into an SMT system, hence unsupervised morphology learning algorithms are more suitable in such cases.

Goldsmith [3] proposes a method of minimum description length (MDL) analysis to model unsupervised learning of the morphology of European languages, attempting to provide both a list of morphemes and an analysis of each word in a corpus. The algorithm described in his work is implemented and named *Linguistica*, and is freely available in [3]. A morphological grammar is developed with the use of a set of heuristics, then MDL is used to determine whether the modifications proposed by the heuristics will be adopted or not, by eliminating inappropriate parses for every word in the corpus. Other work in this area of automatic morphology acquisition include Jacquemin's work [4], Schone and Jurafsky [5], and Baroni et al. [6]. In the next section, a brief introduction into the Linguistica system is provided.

2.1. Linguistica Automatic Morphology Acquisition System

In our work, we used the Linguistica system to perform morphological analysis for both the source and target languages. As mentioned in the previous section, Linguistica uses a set of heuristics to provide an initial morphological analysis. The first one (called take-all-splits), considers for each word of length l all the possible cuts into $w_{1,i}$, $w_{i+1,l}$, $1 \leq i < l$. For each cut, the metric H is computed (as seen in Eq. (1)) and the corresponding probability of the cut is given by Eq. (2), i.e.,

$$H(w_{1,i}, w_{i+1,l}) =$$

$$= -(i \log \text{freq}(\text{stem} = w_{1,i}) +$$

$$(l-i) \log \text{freq}(\text{suffix} = w_{i+1,l})) \qquad (1)$$

where freq represents the number of times a stem or suffix appears in the corpus and

$$prob(w = w_{1,i}\ w_{i+1,l}) = \frac{1}{Z} e^{-H(w_{1,i}, w_{i+1,l})} \qquad (2)$$

where the normalization factor Z equals

$$Z = \sum_{i=1}^{n-1} H(w_{1,i}, w_{i+1,l})$$

For each word, the best parse in the maximum likelihood sense is selected to bootstrap the heuristic and then the metric is optimized globally over all words, stems and suffixes in the corpus (usually the process converges after five iterations).

The second heuristic computes the counts of all sequences of characters with length n between two and six letters. Then for each n-gram $n_1, n_2 \ldots n_k$ we compute the weighted mutual information metric:

$$\frac{[n_1, n_2, \ldots, n_k]}{\text{Total count of } n\text{-grams}} log \frac{[n_1, n_2, \ldots, n_k]}{[n_1][n_2] \ldots [n_k]}$$

The top 100 scoring n-grams are kept and used to parse each word (if possible) into stem plus suffix. For those words that more than one splits exist, the previous heuristic is used to choose the best one.

Finally, for each stem the list of all corresponding suffixes is created which is referred to as a *signature*. Stems with the same suffix signatures are merged. Signatures that contain more than one stems and affixes are referred to as *regular signatures* and are of the form

$$\left\{ \begin{array}{c} \text{stem}_1 \\ \text{stem}_2 \\ \text{stem}_3 \end{array} \right\} \left\{ \begin{array}{c} \text{suffix}_1 \\ \text{suffix}_2 \end{array} \right\}$$

Heuristic rules are used to add stems or suffixes to regular signatures (based on similarities with other regular signatures) thus improving on the generalization power of the morphological rules. Note that the morphological signatures are derived in a fully unsupervised fashion. In Section 4, additional heuristic rules that have been used to improve the Linguistica stemmer performance are presented.

3. INCORPORATING MORPHOLOGY INTO SMT SYSTEMS

In statistical machine translation, the translation problem is posed as a posterior probability maximization problem. If we consider W_s and W_t to be word sequences for the source and target languages respectively, then the problem can be formulated as:

$$\hat{W}_t = \arg\max_{W_t} P(W_t | W_s) \qquad (3)$$

where $\hat{W}_t$ is the translated sequence of words in the target language.

Using the algorithms described in Section 2, we come up with knowledge about the morphology of both the source and target languages. This knowledge can be represented as a (deterministic or statistical) mapping from a sequence of words W to a sequence of stems S. These stems may be extracted in general by the use of a statistical morphological analyzer (stemmer) that computes the probabilities $P(S|W)$. Also, a morphological generator is defined as the model that computes the reverse probabilities $P(W|S)$.

Let us consider S_s and S_t to be sequences of stems for the source and target languages respectively. Then a stem-to-stem machine translation system can be formulated as:

$$\hat{S}_t = \arg\max_{S_t} P(S_t|S_s) \quad (4)$$

Using the statistical models for the morphological analyzer $P(S_s|W_s)$ and morphological generator $P(W_t|S_t)$ for the source and target languages respectively, as well as the stem-to-stem translation model $P(S_t|S_s)$ we may write

$$\begin{aligned}\hat{W}_t &= \arg\max P(W_t|W_s) \\ &= \arg\max_{W_t} \sum_{S_t,S_s} P(W_t, S_t, S_s|W_s) \\ &= \arg\max_{W_t} \sum_{S_t,S_s} P(W_t|S_t, S_s, W_s)P(S_t|S_s, W_s)P(S_s|W_s) \\ &= \arg\max_{W_t} \sum_{S_t,S_s} P(W_t|S_t)P(S_t|S_s)P(S_s|W_s) \quad (5)\end{aligned}$$

provided that W_t, S_s are conditionally independent given S_t; W_t, W_s are conditionally independent given S_t, S_s; and W_s,S_t are conditionally independent given S_s. This equation corresponds to a word-to-word translation model; however, in this system word to word translation is performed via the stem to stem system, i.e., $W_s \to S_s \to S_t \to W_t$.

Eq. (5) can be further simplified as follows: the mapping $S \to W$ is a many to one mapping and $P(S_s|W_s) = 1$, because the mapping $W_s \to S_s$ is deterministic, so the double summation at Eq. (5) becomes a single summation over S_t only, as follows:

$$\hat{W}_t = \arg\max_{W_t} \sum_{S_t} P(W_t|S_t)P(S_t|S_s) \quad (6)$$

We refer to this system as the morphological or stem-based SMT system.

3.1. SMT system combination

Once we have built the morphological SMT system, we need to combine it with the traditional lexical SMT system. This combination can be done by assuming that each SMT system computes probabilities independently of each other, i.e.,

$$\hat{W}_t = \arg\max_{W_t} [P(W_t|W_s)]^{w_0} [\sum_{S_t} P(W_t|S_t)P(S_t|S_s)]^{w_1} \quad (7)$$

where w_0 and w_1 are weights that model the "confidence" we have in each translation, the lexical and the morphological SMT models. By combining these two SMT systems we hope that we overcome the weakness of the conditional independence assumptions of Eq. (5). This combination may be implemented at an early or at a late stage (e.g., word lattice combination).

4. COMBINED LEXICAL-MORPHOLOGICAL SYSTEM IMPLEMENTATION

The combined lexical and morphological SMT system is implemented using late integration and lattice re-scoring according to the following steps:

1. The lexical SMT system that computes the probabilities $P(W_t|W_s)$ is built.
2. The training corpus is stemmed using the unsupervised rules derived from the Linguistica system.
3. The stemmed corpus is used to derive the morphological (stem) SMT system that computes the probabilities $P(S_t|S_s)$.
4. Every sentence in the evaluation corpus is decoded using the lexical SMT system producing a lattice of possible word-level translations. This lattice is then represented as a finite state acceptor F_W.
5. Every sentence in the evaluation corpus is stemmed and then decoded using the morphological SMT system. The resulting lattice contains all possible stem-level translations and is represented as a finite state acceptor F_S.
6. The stem to word model $P(W_t|S_t)$ in the target language is constructed by running the Linguistica system on the target language corpus and obtaining the morphological signature. The stem to word model is represented as a unweighted (costless) finite state transducer T_{SW}, i.e., in our case, we assume that all possible words that can be generated from a stem are equiprobable [1].
7. The stem acceptor F_S and the stem to word transducer T_{SW} are composed to obtain a stem to word

[1] In order to guarantee non-empty composition in the next step, all words contained in F_W and F_S were added as identity mappings in T_{SW} and then Kleene closure was applied to T_{SW}.

mapping; the resulting transducer is projected to its output symbols to obtain the finite state acceptor $F_{W'}$.

8. F_W and $F_{W'}$ acceptors are re-weighted (weights multiplied) by the factors w_0 and w_1 as discussed in Section 3.1. (in practice, we don't weight F_W and w_0 is always 1.).

9. The weighted acceptors F_W and $F_{W'}$ are intersected and the best path of the intersection is found using Viterbi decoding. The best path T' represents the translated sentence of the combined lexical-morphological SMT system.

The process that has been described above can be formulated as follows:

$$T' = \text{bestpath}\{([F_S \circ T_{SW}]_2 * w_1) \cap F_W\}$$

where $\circ$ represents composition, $\cap$ intersection, $*$ weighting and $_2$ projection to the output symbols; T', F_S, T_{SW}, F_W and w_1 are defined above. In our work, we used the AT&T FSM Library [7] for the representation of finite state machines and the operations applied to them (closure, composition, intersection, best path decoding).

5. EXPERIMENTS

Two machine translation systems were developed: the baseline lexical SMT system and the combined lexical-morphological SMT system (according to the implementation plan above). Henceforth we will refer to the lexical SMT system as system A and the lexical-morphological SMT system as system B.

5.1. Corpora

Both systems have been trained on parts of the Europarl corpus [8], a parallel corpus in 11 European languages which is extracted from the proceedings of the European Parliament. Since the size of the corpus that has been used for training affects the translation quality of the systems, we chose to use two different sizes, of 1M and 4M words each. We used the rest of the Europarl documents for development and testing sets. We chose to use English as the source language and Greek as the target language.

Care was taken when creating training and test sets: it is common for sequential segments of text in the corpus to share the same vocabulary and style, so it is better to avoid creating models based on such data. In order to overcome this, we used broad sampling, so the chosen data is evenly distributed in the corpus.

5.2. System A: Lexical SMT

For our simple lexical SMT system, we used the GIZA++ system [9] to obtain alignment on the training corpus and then we trained phrase-based statistical machine translation models, as well as a language model for the target language. Then, given a sentence in the source language, we use the Pharaoh decoder [10] to compute the best translation for this sentence, using the models discussed.

5.3. System B: Lexical-Morphological SMT

For the purpose of our morphological SMT system, morphological analysis is performed on W_s, stems are extracted for S_s and affixes are ignored. The same process is performed for the target language. In order to do this, we had to derive morphological information about both languages in an either supervised or unsupervised way.

We used the Linguistica morphological analyzer to automatically derive morphological rules from a 5M word parallel translation corpus, in both the source and target languages, in an unsupervised way. As we have discussed in Section 2.1, Linguistica uses a set of heuristics to develop a probabilistic grammar and then depends on Minimum Description Length analysis to determine which of the rules proposed will be adopted. For a 5M word English corpus precision of 85.9% and recall of 90.4% are reported.

In order to increase precision, at the expense of recall, two additional heuristics are used; each word analyzed by Linguistica is considered for stemming only if it lays above a given length d. In addition to this, the ratio r of the length of the suffix per whole word length must not exceed a threshold. The values used were d=6 and r=0.3. For a 2k set of distinct Greek words, Linguistica scored a 79.5% precision, while after the incorporation of the heuristics just described, precision reached 93.8%.

Using the morphological rules obtained by Linguistica, the parallel corpus was stemmed and phrase-based models were trained, in the same fashion as with the lexical SMT system. A language model for the target language, based on the stemmed corpus was constructed as well. The training data used was exactly the same as the data that has been used to train the corresponding models of the lexical SMT. The stem to word model was constructed and the combination of the word and stem-based system was performed using finite state machines, late integration and lattice combination as outlined in Section 4.

6. RESULTS

We evaluated our systems output using the BLEU [11] and NIST [12] evaluation metrics. We provided both systems with an initial test data of approximately 26k sentences. For the 1M corpus, system B yielded approximately 11k sentences that were different compared to those of system A. For the 4M corpus, the different sentences were found to be about 6k. This number of different sentences, is function of the size of the lattices used by the SMT system, thus enlarging these lattices could produce more different translations.

Table 1 summarizes this information; for every training corpus size (TC_s), the number of different sentences that formed the evaluation set are displayed (Ds), as well as the ratio (Ds_r) of the different sentences compared to the total 26k sentences of the test set.

TC_s	Ds	Ds_r
1M	10669	40.57%
4M	6322	24.04%

Table 1. Training and evaluation set sizes

For every training corpus size, we performed several experiments by changing the weight w_1 of the FSM containing the morphological information. In order to focus on the real improvement of the new system, our evaluation set does not consist of all the sentences that form the test data, but those that resulted in different translations between the two systems. The results of these experiments are shown in Tables 2 and 3.

TC_s	w_1	N_A	N_B	N_B-N_A	Improvem.
1M	0.05	3.0253	3.4579	0.4326	14.30%
1M	0.1	3.0538	3.4618	0.4080	13.36%
1M	0.2	3.1144	3.4663	0.3519	11.30%
1M	0.3	3.1539	3.4453	0.2914	9.24%
4M	0.1	4.2391	4.4139	0.1748	4.12%

Table 2. NIST scores for systems A and B for various combination weights and training corpus sizes

TC_S	w_1	B_A	B_B	B_B-B_A	Improvem.
1M	0.05	0.0604	0.0693	0.0089	14.74%
1M	0.1	0.0611	0.0697	0.0086	14.08%
1M	0.2	0.0629	0.0704	0.0075	11.92%
1M	0.3	0.0635	0.0703	0.0068	10.71%
4M	0.1	0.1006	0.1057	0.0051	5.07%

Table 3. BLEU scores for systems A and B for various combination weights and training corpus sizes

The best score improvement achieved was 0.4326 for the NIST scores and 0.0089 for the BLEU evaluation metrics, which correspond to 14.30% and 14.74% relative score improvement respectively compared to the lexical system. The tables show that smaller weights w_1 provide bigger improvements, i.e., the FSM containing the morphological information should be weighted more. This is probably due to the fact that the statistics of the word stems are better trained than the statistics of the words for training sets of the same size, i.e., the stem model is better trained than the word model. It can also be seen that the incorporation of morphological information provides more improvement for systems that have been trained with smaller data sets, since the scores for the 1M training corpus are much better than those of the 4M training corpus.

In order to be certain that our test set size is large enough to guarantee true improvement, we perform bootstrap resampling [13]. This method has been used in various fields of research, including automatic speech recognition and statistical machine translation [14], [15], [16].

TC_s	w_1	N_d mean	N_d interval	N_B RSD
1M	0.05	0.4325	[0.3996, 0.4654]	0.77%
1M	0.1	0.4082	[0.3751, 0.4405]	0.75%
1M	0.2	0.3520	[0.3199, 0.3836]	0.75%
1M	0.3	0.2913	[0.2600, 0,3217]	0.76%
4M	0.1	0.1748	[0.1298, 0.2204]	0.84%

Table 4. 95% confidence intervals for N_d scores (NIST)

Table 4 shows the interval mean and the 95% confidence interval for the differences in the NIST scores between systems A and B ($N_d = N_B - N_A$). Assuming the bootstrap hypothesis, we can say that there is 95% confidence that, for example, for the system with $w = 0.01$ the improvement of the NIST score lies between 0.4220 and 0.4887, which corresponds to improvement between 14.06% and 16.28%. The last column of this table also shows the relative standard deviation for the NIST value of the morphological SMT system (*Relative Standard Deviation* or RSD is defined as $(100 * \sigma/\mu)\%$, where μ and σ are the mean and standard deviation respectively). Clearly, the improvements achieved by combining the lexical and the morphological information is statistically significant.

7. CONCLUSION AND FUTURE WORK

In this paper, we described a method for the incorporation of morphological knowledge into statistical machine translation (SMT) systems. We proposed an implementation of a stem-based SMT system and its combination with traditional lexical SMT systems. Experiments showed statistically significant improvements of this combined system when compared to a purely lexical one. The combined lexical-morphological SMT system improves on the baseline system for both the NIST and BLEU evaluation metrics.

Currently our system ignores affix information. The lexical-morphological system could be enhanced by modeling the probabilities of the sequences of affixes (as well as that of stems) and combining the affix model into the SMT formulation. Also, in the future we are looking forward into conducting experiments with larger development and test data sets as well as different language pairs, in order to fine tune our system and provide further evidence on the improvement that morphological (and in general linguistic) information can provide in statistical machine translation systems.

Acknowledgments

We would like to thank Fanouris Moraitis, for developing the baseline lexical SMT system on which our research was based. This work was partially supported by the EU-IST-FP6 MUSCLE network of excellence.

8. REFERENCES

[1] U. Germann, "*Building a Statistical Machine Translation System from Scratch: How Much Bang for the Buck Can We Expect?*," ACL 2001 Workshop on Data-Driven Machine Translation, Toulouse, France, July 7, 2001.

[2] S. Niessen, H. Ney, "*Toward hierarchical models for statistical machine translation of inflected languages*," In: Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation, 2001.

[3] J. Goldsmith, "*Unsupervised Learning of the Morphology of a Natural Language*," In: Computational Linguistics, Volume 27, Number 2, 2001.

[4] C. Jacquemin, "*Guessing Morphology from Terms and Corpora*," Research and Development in Information Retrieval, pp. 156-165., 1997.

[5] P. Schone and D. Jurafsky, "*Knowlege-Free Induction of Inflectional Morphologies*," In: Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL-2001), 2001.

[6] M. Baroni, J. Matiasek, H. Trost, "*Unsupervised discovery of morphologically related words based on orthographic and semantic similarity*," In: Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), Association for Computational Linguistics, pp. 48-57., Philadelphia, jul 2002.

[7] M. Mohri, F. C. N. Pereira, M. D. Riley, "*AT&T FSM Library TM - Finite-State Machine Library*," .

[8] P. Koehn, "*Europarl: A Multilingual Corpus for Evaluation of Machine Translation*," Draft, Unpublished.

[9] F. J. Och, H. Ney, "*A Systematic Comparison of Various Statistical Alignment Models*," Computational Linguistics, volume 29, number 1, pp. 19-51, mar 2003.

[10] P. Koehn, "*Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models*," The 6th Conference of the Association for Machine Translation in the Americas (AMTA), 2004.

[11] K. Papineni and S. Roukos and T. Ward and W. Zhu, "*Bleu: A Method for Automatic Evaluation of Machine Translation*," Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

[12] K. Papineni and S. Roukos and T. Ward and W. Zhu, "*Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*," .

[13] B. Efron, and R.J. Tibshirani, "*An Introduction to the Bootstrap*," CRC Press, 1994.

[14] M. Bisani and H. Ney, "*Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation*," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), may 2004.

[15] Y. Zhang, S. Vogel, "*Measuring Confidence Intervals for the Machine Translation Evaluation Metrics*," In: Proceedings of The 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2004), Baltimore, MD USA, oct 2004.

[16] Y. Zhang, S. Vogel, A. Waibel, "*Interpreting BLEU/NIST scores: How much improvement do we need to have a better system?*," In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, may 2004.