**RC25274 (WAT1203-074) March 28, 2012**
**Computer Science**

# IBM Research Report

## An Auxiliary Phrase Table Approach to Closed-Loop Multi-Pass SMT

**Juan M. Huerta**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

# An Auxiliary Phrase Table Approach to Closed-Loop Multi-Pass SMT

**Juan M. Huerta**
IBM T J Watson Research Center
1101 Kitchawan Road
Yorktown Heights USA
huerta@us.ibm.com

## Abstract

We describe a new approach to SMT based on closed loop multi-pass decoding as well as on the use of auxiliary phrase tables. In our approach, portions of an initial translation output are automatically selected, matched with their input segments, modified under specific criteria, and reintroduced to subsequent translation passes in the form of phrase tables. The motivation behind this approach is that the generation of rich morphological output (e.g., gender, person, tense) is a problem not easily resolvable within a single decoding iteration but rather, is better addressed after the output of an initial translation has been established. Our SMT experiments show consistent BLEU score improvements under several configurations in an eSupport domain translation test.

## 1 Introduction

In this work we propose a general closed loop multi-pass approach to Statistical Machine Translation. In our approach, the output of individual translation passes is analyzed and used to dynamically generate phrase tables which are used in subsequence SMT iterations.

While our approach is generalizable to multiple translation passes and can address a broad array of issues, in this paper we specifically focus on a 2-pass configuration and on modifying particular syntactic arrangements. We focus also on the problem of translating in the direction that increases morphological richness. For example, when translating from English into Spanish, nouns and adjectives acquire gender information which is not contained in the English sentence.

The organization is as follows: we first explain the basic idea behind our approach, followed by an explanation of how our work relates to other approaches. We then describe the specific scope and issues we presently focus on, followed by a description of our experimental setup and results. We conclude with a brief discussion.

## 2 Approach

Figure 1 shows a diagram of the proposed closed loop multi-pass strategy in which the output of each SMT iteration is analyzed and used in subsequent decoding passes in the form of phrase tables (which we call auxiliary phrase tables).

The analysis we propose can be organized into two categories: processing focused on sentence context, and processing focused on domain context. For the sentence context we propose the following steps (Section 4 provides more detail on the particular dependency types and morphology issues we focus on, as well as detailed examples):

- Identify instances of specific relation or link types in the input sentence using a dependency parser in the source language (English, in our case). Identify pairs of words associated with these links. Each link has a *head* word and a *dependent* word.
- Identify word and phrase regions in the SMT output which are associated with each of the *dependent* words in the links identified. For this we rely on SMT decoding information as well as on

dependency information. Altogether these constitute the candidate phrases.

- Label the SMT output using a gender, person and conjugation sensitive POS tagger in the target language (Spanish, in our case).
- Using POS information, identify inflectional discrepancies for each candidate phrase between the portion of the translated output corresponding to the head of the relation associated with a given phrase and the dependent's word portion translation.
- For those candidate phrases in which a discrepancy has been observed, modify the target portion of the phrase using an inflector and introduce the result as a phrase in the auxiliary phrase table. Make available the annotated input to subsequent iterations as a way to associate specific phrases with their corresponding sentences.

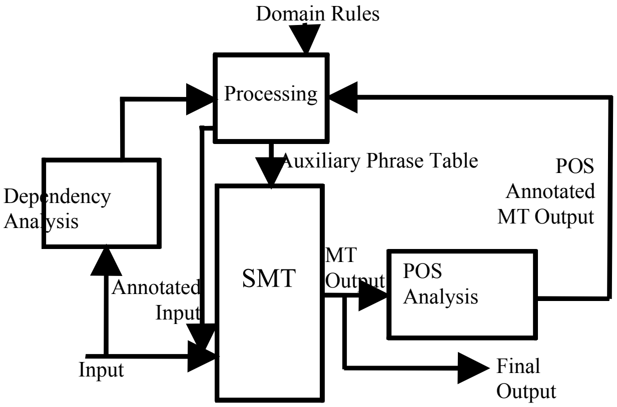

**Figure 1.** General multi-pass SMT system

It can be argued that a phrase-based SMT decoder using an SML of reasonable order should be able to resolve most of the gender, number and conjugation discrepancies in a typical single-pass decoding strategy. This should be the case because the context influencing the morphology of translated words is typically located within sufficiently close distance and this should allow the LM to resolve these issues successfully. However, in practice, it is not unusual for the Language Model span to be of insufficient order for successfully resolving this type of issues. Furthermore, the sentence to be translated might contain nested constructs (e.g., noun phrases, etc.) that introduce a separation between affected words and their defining words. Finally, rich morphology can cause phrase explosion which combined with model pruning might result in incomplete models. Therefore, we believe it is advantageous to explicitly address inflectional and conjugational agreement using the multi-iteration strategy we are proposing.

## 3 Related Work

Multi-pass MT has been previously proposed in the context of tree based translation (Petrov et al. (2008)) using coarse-to-fine language model complexity with the intention of addressing computational cost. Another multi-decoding approach was proposed by Chen et al. (2008) in which multiple systems are run in parallel, followed by a combination and rescoring of their outputs. Chiang (2007) proposed a hierarchical phrase based approach to deal with rich morphological output. Our approach differs from all these prior approaches by following a *closed loop* phrase based multi-pass strategy in which SMT outputs and inputs are analyzed and used to dynamically extend the translation models used in subsequent passes. The decoding algorithm is essentially the same across decoding passes.

On the other hand, MT output post-editing (e.g., Simard et al. (2007)) can be seen as a two pass system in which the output of MT is adjusted under static rules. In contrast to our method, this approach does not constitute a closed loop approach as no subsequent decoding passes are carried out.

## 4 Processing Details

As previously mentioned, the processing we perform is of 2 types: processing that focuses on the sentence context, and processing that focuses on the domain context.

### 4.1 Sentence Context Processing

We provide an example in order to describe this type of processing. In our example, we want to translate the following English sentence:

*Service pack point releases will be supported.*

Using the a dependency parser (e.g., de Marmeffe et al.) we generated typed dependencies and focused particularly in the `nsubjpass` dependency (passive nominal subject) which we know typically generates constructions in Spanish that requires gender and number resolution. The relation of interest is, in our particular example,

*nsubjpass(supported-7, releases-4)*

In this case the head word is *supported* and the dependent is *releases*. Using an SMT system we produce the first translation output which we label using a inflection sensitive POS tagger (below are shown only noun and adjective tags (NN, ADJ) with gender (m/f) and number (s/p) information):

*Punto los releases NN:pm de servicio NN:sm del paquete NN:sm de estará NN:sm soportado ADJ:sm*

Using information from the decoder we associate the dependent English word of interest (*releases*) with its Spanish counterpart (*releases*) (which has POS label NN:pm). Similarly we associate *soportado* (with POS label ADJ:sm) with *supported* which is the head word in the dependency relation under analysis. This brings together *releases* with *soportado* but most importantly, it makes evident that the number in the NN (plural) tag is in disagreement with the number in the ADJ (singular) tag. Therefore, it is necessary to adjust the number (but not the gender) of *soportado* resulting in *soportados*. This can be done using a simple rule based inflector. The phrase *supported* → *soportados* is then ready to be introduced into the auxiliary phrase table. However, this sentence might already exist in the SMT's phrase table: to ensure that this particular sentence is actually used we increase the weight of this phrase (i.e., we are essentially forcing its use in subsequent passes).

In order to ensure that the correct instance of the word *supported* is associated with the newly created phrase (in case there were more than one instances of the word *supported* in the sentence or in the corpus translated) we append a unique identifier to the word *supported* both in the phrase table as well as in the input string and thus create a new annotated input. The input string becomes:

*Service pack point releases will be supported-01234.*

And the auxiliary translation table now includes the phrase:

*supported-01234* → *soportados*

It should be noted that the addition of the identifier string in the input sentence as well as in the auxiliary phrase table does not interfere with the normal use of the target language SLM during translation.

### 4.2 Domain Context Processing

The processing described above focuses on creating new phrases that better suit the context of each translation sentence. Sometimes, it is necessary to perform adjustments to translation output to better suit particular domain translation styles. For example, a system that has been trained on substantial amounts of news and similar material will naturally be biased towards conjugations that are common in news but less so in technical domains.

For example, the sentence:

*The final update for this information was published in online softcopy form in August 2008*

Has as its first translation:

*La actualización final que se ha publicado para esta información en formato de copia software en línea en agosto de 2008*

While the phrase *was published* → *se ha publicado* is reasonable, it might be desirable, for this domain's stylistic purposes to avoid the impersonal conjugation and use instead the passive voice in the Spanish pretérito tense: *was published* → *fue publicada*. Thus, our approach will pick an existing conjugated instance, identify its infinitive (*publicar*), and generate new inflections. In this particular case we can introduce 2 phrases (one for the masculine and one for the feminine) and allow the decoder to pick the gender based on each sentence's defining noun (*actualización*, in our example) which is feminine in this case. In our particular example, the second pass translation becomes:

*La actualización final que fue publicada para esta información en formato de copia software en línea en agosto de 2008*

Because this processing is addressing largely domain-specific issues (rather than sentence specific), it is not necessary to provide unique identifiers in the phrase table or input sentences but rather allow the SMT to pick the correct gender.

## 5 Experiments

Our experiments were based on an English-to-Spanish phrase-based SMT system trained on 26M general-domain sentence pairs. The resulting phrase model is composed of 40M phrases. The SLM is a smoothed 5-gram backoff statistical LM trained on approximately 600M general domain sentences. We used the Stanford Dependency Parser (Marneffe et al. (2006)) and a Spanish MaxEnt POS tagger using gender, number and conjugation sensitive tags with 94.6% accuracy.

We used two test sets: a development set and an evaluation set both in the eSupport domain. The eSupport domain generally consists of sentences originating from manuals, web pages and documents intended to support the end user in products related to computer software, computer hardware, and related services. The Development set consists of 600 sentences (10.4K words) while the Evaluation set consists of 1038 sentences (19.6K words).

We focused on experimenting under the following configurations: baseline, blind post editing, domain context processing (in various conditions: single gender, multiple genders, matched pre-computed table, and mismatched pre-computed table), and sentence context processing. The system was limited to a 2-pass configuration in which the first pass is used to generate the auxiliary phrase tables which are applied in a second translation pass. Table 1 shows the results obtained.

We measured the BLEUr1n4c (1 reference, 4 gram, case and punctuation sensitive) and the baseline for the Development set and the Evaluation set was 35.45 and 36.37, respectively.

| | **Development BLEUr1n4c** | **Evaluation BLEUr1n4c** |
|---|---|---|
| Baseline | 35.45 | 36.67 |
| Blind Post Editing | 35.35 | 36.39 |
| Domain Context (single gender) | 38.61 | 36.91 |
| Domain Context (multiple gender) | 38.76 | 36.94 |
| Sentence Context | 36.26 | 37.20 |
| Sentence Context + Domain Context | 37.52 | 37.36 |

**Table 1**. Experimental Results (BLEU scores)

We measured the impact of post editing the SMT baseline output using domain context rules (i.e., passive voice focus). This slightly decreased the translation performance (Table 1 line 2).

In the following experiment (Table 1 line 3) we introduced domain context modifications (Section 4.2) but allowed only one arbitrary gender per rule thus relaxing the need of a gender analysis step. Furthermore, we evaluated the impact of mismatched conditions on the test set by using only the Development set to create the auxiliary phrase table. While the performance for the Development set improved substantially, the increase is not so substantial for the Evaluation set.

We repeated the previous experiment but introducing both genders for each newly created phrase and allowing the SMT to decide which phrase to use (Table 1 line 4). The BLEU score further improved on both sets.

We then performed sentence context experiments following the process described in Section 4.1 (Table 1 line 5). We did not do this under mismatched conditions, but instead, we allowed separate processing for each sentence in the Development and Evaluation sets. We observed improvements on both sets over their respective baselines and observed that the Evaluation set obtains a larger gain than under the mismatched domain context experiments.

In the final configuration (Table 1 line 6), we combined both types of auxiliary phrase tables (sentence & domain context) but contextual phrases were weighted down using an arbitrary constant weight (i.e., not given the maximum possible weight). This was done to illustrate the effect of forcing sentence context phrases while providing domain context phrases with less weight. This configuration resulted in the best scores for the Evaluation set: 0.7 BLEU points over the baseline.

## 6 Discussion

We introduced a new approach to SMT based on a multi-pass approach using auxiliary phrase tables. In our experiments, the best configuration for the Development set used exclusively domain context auxiliary phrase tables. As similar large gains were not observed under mismatched conditions suggest the importance of dynamically creating auxiliary phrase tables matching particular data sets. The best configuration for the Evaluation set used sentence context auxiliary phrase tables enhanced with (mismatched) domain context phrases, which lags behind the gain obtained in matched conditions. This suggests that a potential area of future improvement is the calculation of optimal weights when mixing domain & sentence context derived phrase tables.

## References

B. Chen, D. Xiong, M. Zhang, A. Aw, and H. Li. 2008. "I2r multi-pass machine translation system for iwslt 2008," in IWSLT, 2008.

D. Chiang. 2007. "Hierarchical phrase-based translation," Computational Linguistics, vol. 33, no. 2, 2007.

M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn. 2007. "Rulebased translation with statistical phrase-based postediting," in 2nd Workshop on SMT 2007.

M. C. de Marneffe, B. MacCartney and C. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

S. Petrov, A. Haghighi, and D. Klein. 2008. Coarse-to-fine syntactic machine translation using language projections. In *EMNLP 2008.*