UCB System Description for the WMT 2007 Shared Task

Preslav Nakov
EECS, CS division
University of California at Berkeley
Berkeley, CA 94720
nakov@cs.berkeley.edu

Marti Hearst
School of Information
University of California at Berkeley
Berkeley, CA 94720
hearst@ischool.berkeley.edu

Abstract

For the WMT 2007 shared task, the UC Berkeley team employed three techniques of interest. First, we used monolingual syntactic paraphrases to provide syntactic variety to the source training set sentences. Second, we trained two language models: a small in-domain model and a large out-of-domain model. Finally, we made use of results from prior research that shows that cognate pairs can improve word alignments. We contributed runs translating English to Spanish, French, and German using various combinations of these techniques.

1 Introduction

Modern Statistical Machine Translation (SMT) systems are trained on aligned sentences of bilingual corpora, typically from one domain. When tested on text from that same domain, such systems demonstrate state-of-the art performance; however, on out-of-domain text the results can get significantly worse. For example, on the WMT 2006 Shared Task evaluation, the French to English translation BLEU scores dropped from about 30 to about 20 for nearly all systems, when tested on *News Commentary* rather than *Europarl* (Koehn and Monz, 2006).

Therefore, this year the shared task organizers have provided 1M words of bilingual *News Commentary* training data in addition to the *Europarl* data (about 30M words), thus challenging the participants to experiment with domain adaptation.

Below we describe our domain adaptation experiments, trying to achieve better results on the *News Commentary* data. In addition to training on both data sets, we make use of monolingual syntactic paraphrases of the English side of the data.

2 Monolingual Syntactic Paraphrasing

In many cases, the testing text contains "phrases" that are equivalent, but syntactically different from the phrases learned on training, and the potential for a high-quality translation is missed. We address this problem by using nearly equivalent syntactic paraphrases of the original sentences. Each paraphrased sentence is paired with the foreign translation that is associated with the original sentence in the training data. This augmented training corpus can then be used to train an SMT system. Alternatively, we can paraphrase the test sentences making them closer to the target language syntax.

Given an English sentence, we parse it with the Stanford parser (Klein and Manning, 2003) and then generate paraphrases using the following syntactic transformations:

1. $[_{\mathbf{NP}}\ \mathbf{NP}_1\ \mathbf{P}\ \mathbf{NP}_2] \Rightarrow [_{\mathbf{NP}}\ \mathbf{NP}_2\ \mathbf{NP}_1]$.
   inequality in income ⇒ *income inequality*.
2. $[_{\mathbf{NP}}\ \mathbf{NP}_1\ \boldsymbol{of}\ \mathbf{NP}_2] \Rightarrow [_{\mathbf{NP}}\ \mathbf{NP}_2\ \mathbf{poss}\ \mathbf{NP}_1]$.
   inequality of income ⇒ *income's inequality*.
3. $\mathbf{NP}_{poss} \Rightarrow \mathbf{NP}$.
   income's inequality ⇒ *income inequality*.
4. $\mathbf{NP}_{poss} \Rightarrow \mathbf{NP}_{PP_{of}}$.
   income's inequality ⇒ *inequality of income*.
5. $\mathbf{NP}_{NC} \Rightarrow \mathbf{NP}_{poss}$.
   income inequality ⇒ *income's inequality*.
6. $\mathbf{NP}_{NC} \Rightarrow \mathbf{NP}_{PP}$.
   income inequality ⇒ *inequality in incomes*.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 212–215,
Prague, June 2007.

Sharply rising income inequality has raised the stakes of the economic game .
Sharply rising *income inequality* has raised the *economic game 's stakes* .
Sharply rising *income inequality* has raised the *economic game stakes* .
Sharply rising *inequality of income* has raised the *stakes of the economic game* .
Sharply rising *inequality of income* has raised the *economic game 's stakes* .
Sharply rising *inequality of income* has raised the *economic game stakes* .
Sharply rising *inequality of incomes* has raised the *stakes of the economic game* .
Sharply rising *inequality of incomes* has raised the *economic game 's stakes* .
Sharply rising *inequality of incomes* has raised the *economic game stakes* .
Sharply rising *inequality in income* has raised the *stakes of the economic game* .
Sharply rising *inequality in income* has raised the *economic game 's stakes* .
Sharply rising *inequality in income* has raised the *economic game stakes* .
Sharply rising *inequality in incomes* has raised the *stakes of the economic game* .
Sharply rising *inequality in incomes* has raised the *economic game 's stakes* .
Sharply rising *inequality in incomes* has raised the *economic game stakes* .

Table 1: Sample sentence and automatically generated paraphrases. Paraphrased NCs are in italics.

7. **remove *that* where optional**
   I think that *he is right* ⇒ *I think he is right*.
8. **add *that* where optional**
   I think he is right ⇒ *I think* that *he is right*.

where:

poss	possessive marker: ' or 's;
P	preposition;
NP$_{PP}$	NP with internal PP-attachment;
NP$_{PP_{of}}$	NP with internal PP headed by *of*;
NP$_{poss}$	NP with internal possessive marker;
NP$_{NC}$	NP that is a Noun Compound.

While the first four and the last two transformations are purely syntactic, (5) and (6) are not. The algorithm must determine whether a possessive marker is feasible for (5) and must choose the correct preposition for (6). In either case, for noun compounds (NCs) of length 3 or more, it also needs to choose the position to modify, e.g., *inquiry's committee chairman* vs. *inquiry committee's chairman*.

In order to ensure accuracy of the paraphrases, we use statistics gathered from the Web, using a variation of the approaches presented in Lapata and Keller (2004) and Nakov and Hearst (2005). We use patterns to generate possible prepositional or copula paraphrases in the context of the preceding and the following word in the sentence, First we split the NC into two parts N_1 and N_2 in all possible ways, e.g., *beef import ban lifting* would be split as: (a) N_1="*beef*", N_2="*import ban lifting*", (b) N_1="*beef import*", N_2="*ban lifting*", and (c) N_1="*beef import ban*", N_2="*lifting*". For every split, we issue exact phrase queries to the Google search engine using the following patterns:

"lt N_1 poss N_2 rt"
"lt N_2 prep det N_1' rt"
"lt N_2 that be det N_1' rt"
"lt N_2 that be prep det N_1' rt"

where: `lt` is the word preceding N_1 in the original sentence or empty if none, `rt` is the word following N_2 in the original sentence or empty if none, `poss` is a possessive marker ('s or '), `that` is *that*, *which* or *who*, `be` is *is* or *are*, `det` is a determiner (*the*, *a*, *an*, or none), `prep` is one of the 8 prepositions used by Lauer (1995) for semantic interpretation of NCs: *about*, *at*, *for*, *from*, *in*, *of*, *on*, and *with*, and N_1' can be either N_1, or N_1 with the number of its last word changed from singular/plural to plural/singular.

For all splits, we collect the number of page hits for each instantiation of each pattern, filtering out the paraphrases whose page hit count is less than 10. We then calculate the total number of page hits H for all paraphrases (for all splits and all patterns), and retain those ones whose page hits count is at least 10% of H. Note that this allows for multiple paraphrases of an NC. If no paraphrases are retained, we repeat the above procedure with `lt` set to the empty string. If there are still no good paraphrases, we set the `rt` to the empty string. If this does not help either, we make a final attempt, by setting both `lt` and `rt` to the empty string.

Table 1 shows the paraphrases for a sample sentence. We can see that *income inequality* is paraphrased as *inequality of income*, *inequality of incomes*, *inequality in income* and *inequality in incomes*; also *economic game's stakes* becomes *economic game stakes* and *stakes of the economic game*.

3 Experiments

Table 2 shows a summary of our submissions: the official runs are marked with a $\star$. For our experiments, we used the baseline system, provided by the organizers, which we modified in different ways, as described below.

3.1 Domain Adaptation

All our systems were trained on both corpora.

- **Language models**. We used two language models (LM) – a small in-domain one (trained on *News Commentary*) and a big out-of-domain one (trained on *Europarl*). For example, for EN $\rightarrow$ ES (from English to Spanish), on the lowercased tuning data set, using in-domain LM only achieved a BLEU of 0.332910, while using both LMs yielded 0.354927, a significant effect.

- **Cognates.** Previous research has found that using cognates can help get better word alignments (and ultimately better MT results), especially in case of a small training set. We used the method described in (Kondrak et al., 2003) in order to extract cognates from the two data sets. We then added them as sentence pairs to the *News Commentary* corpus before training the word alignment models[1] for *ucb3*, *ucb4* and *ucb5*.

- **Phrases.** The *ucb5* system uses the *Europarl* data in order to learn an additional phrase table and an additional lexicalized re-ordering model.

3.2 Paraphrasing the Training Set

In two of our experiments (*ucb3*, *ucb4* and *ucb5*), we used a paraphrased version of the training *News Commentary* data, using all rules (1)-(8). We trained two separate MT systems: one on the original corpus, and another one on the paraphrased version. We then used both resulting lexicalized re-ordering models and a merged phrase table with extra parameters: if a phrase appeared in both phrase tables, it now had 9 instead of 5 parameters (4 from each table, plus a phrase penalty), and if it was in one of the phrase tables only, the 4 missing parameters were filled with 1e-40.

The *ucb5* system is also trained on *Europarl*, yielding a third lexicalized re-ordering model and adding 4 new parameters to the phrase table entries.

Unfortunately, longer sentences (up to 100 tokens, rather than 40), longer phrases (up to 10 tokens, rather than 7), two LMs (rather than just one), higher-order LMs (order 7, rather than 3), multiple higher-order lexicalized re-ordering models (up to 3), etc. all contributed to increased system's complexity, and, as a result, time limitations prevented us from performing minimum-error-rate training (MERT) (Och, 2003) for *ucb3*, *ucb4* and *ucb5*. Therefore, we used the MERT parameter values from *ucb1* instead, e.g. the first 4 phrase weights of *ucb1* were divided by two, copied twice and used in *ucb3* as the first 8 phrase-table parameters. The extra 4 parameters of *ucb5* came from training a separate MT system on the *Europarl* data (scaled accordingly).

3.3 Paraphrasing the Test Set

In some of our experiments (*ucb2* and *ucb4*), given a test sentence, we generated the single most-likely paraphrase, which makes it syntactically closer to Spanish and French. Unlike English, which makes extensive use of noun compounds, these languages strongly prefer connecting the nouns with a preposition (and less often turning a noun into an adjective). Therefore, we paraphrased all NCs using prepositions, by applying rules (4) and (6). In addition, we

[1]Following (Kondrak et al., 2003), we considered words of length 4 or more, we required the length ratio to be between $\frac{7}{10}$ and $\frac{10}{7}$, and we accepted as potential cognates all pairs for which the longest common subsequence ratio (LCSR) was 0.58 or more. We repeated 3 times the cognate pairs extracted from the *Europarl*, and 4 times the ones from *News Commentary*.

Languages	System	LM size		Paraphrasing		Cognates?	Extra phrases	MERT
		News	*Europarl*	*train?*	*test?*		*Europarl*	*finished?*
EN → ES	ucb1⋆	3	5					+
	ucb2	3	5		+			+
	ucb3	5	7	+		+		
	ucb4	5	7	+	+	+		
	ucb5	5	7	+		+	+	
EN → FR	ucb3	5	7	+		+		
	ucb4⋆	5	7	+	+	+		
EN → DE	ucb1⋆	5	7			+		+
	ucb2	5	7		+	+		+

Table 2: **Summary of our submissions.** All runs are for the *News Commentary* test data. The official submissions are marked with a star.

applied rule (8), since its Spanish/French equivalent *que* (as well as the German *daß*) is always obligatory. These transformations affected 927 out of the 2007 test sentences. We also used this transformed data set when translating to German (however, German uses NCs as much as English does).

3.4 Other Non-standard Settings

Below we discuss some non-standard settings that differ from the ones suggested by the organizers in their baseline system. First, following Birch et al. (2006), who found that higher-order LMs give better results[2], we used a 5-gram LM for *News Commentary*, and 7-gram LM for *Europarl* (as opposed to 3-gram, as done normally). Second, for all runs we trained our systems on all sentences of length up to 100 (rather than 40, as suggested in the baseline system). Third, we used a maximum phrase length limit of 10 (rather than 7, as typically done). Fourth, we used *both* a lexicalized and distance-based reordering models (as opposed to lexicalized only, as in the baseline system). Finally, while we did not use any resources other than the ones provided by the shared task organizers, we made use of Web frequencies when paraphrasing the training corpus, as explained above.

4 Conclusions and Future Work

We have presented various approaches to domain adaptation and their combinations. Unfortunately, computational complexity and time limitations prevented us from doing proper MERT for the interesting more complex systems. We plan to do a proper MERT training and to study the impact of the individual components in isolation.

Acknowledgements: This work supported in part by NSF DBI-0317510.

[2]They used a 5-gram LM trained on *Europarl*, but we pushed the idea further, using a 7-gram LM with a Kneser-Ney smoothing.

References

Alexandra Birch, Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Constraining the phrase-based, joint probability statistical translation model. In *Proc. of Workshop on Statistical Machine Translation*, pages 154–157.

Dan Klein and Christopher Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL '03*.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of NAACL*, pages 46–48.

Mirella Lapata and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of nlp tasks. In *Proceedings of HLT-NAACL '04*.

Mark Lauer. 1995. Corpus statistics meet the noun compound: some empirical results. In *Proceedings of ACL '95*.

Preslav Nakov and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of CoNLL '05*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.