LIMSI's statistical translation systems for WMT'10

Alexandre Allauzen, Josep M. Crego, İlknur Durgar El-Kahlout and François Yvon
LIMSI/CNRS and Université Paris-Sud 11, France
BP 133, 91403 Orsay Cedex
Firstname.Lastname@limsi.fr

Abstract

This paper describes our Statistical Machine Translation systems for the WMT10 evaluation, where LIMSI participated for two language pairs (French-English and German-English, in both directions). For German-English, we concentrated on normalizing the German side through a proper preprocessing, aimed at reducing the lexical redundancy and at splitting complex compounds. For French-English, we studied two extensions of our in-house N-*code* decoder: firstly, the effect of integrating a new bilingual reordering model; second, the use of adaptation techniques for the translation model. For both set of experiments, we report the improvements obtained on the development and test data.

1 Introduction

LIMSI took part in the WMT 2010 evaluation campaign and developed systems for two languages pairs: French-English and German-English in both directions. For German-English, we focused on preprocessing issues and performed a series of experiments aimed at normalizing the German side by removing some of the lexical redundancy and by splitting compounds. For this pair, all the experiments were performed using the Moses decoder (Koehn et al., 2007). For French-English, we studied two extensions of our n-gram based system: first, the effect of integrating a new bilingual reordering model; second, the use of adaptation techniques for the translation model. Decoding is performed using our in-house N-*code* (Mariño et al., 2006) decoder.

2 System architecture and resources

In this section, we describe the main characteristics of the phrase-based systems developed for this evaluation and the resources that were used to train our models. As far as resources go, we used all the data supplied by the 2010 evaluation organizers. Based on our previous experiments (Déchelotte et al., 2008) which have demonstrated that better normalization tools provide better *BLEU* scores (Papineni et *al.*, 2002), we took advantage of our in-house text processing tools for the tokenization and detokenization steps. Only for German data did we used the TreeTagger (Schmid, 1994) tokenizer. Similar to last year's experiments, all of our systems are built in "true-case".

3 German-English systems

As German is morphologically more complex than English, the default policy which consists in treating each word form independently from the others is plagued with data sparsity, which poses a number of difficulties both at training and decoding time. When aligning parallel texts at the word level, German compound words typically tend to align with more than one English word; this, in turn, tends to increase the number of possible translation counterparts for each English type, and to make the corresponding alignment scores less reliable. In decoding, new compounds or unseen morphological variants of existing words artificially increase the number out-of-vocabulary (OOV) forms, which severely hurts the overall translation quality. Several researchers have proposed normalization (Niessen and Ney, 2004; Corston-oliver and Gamon, 2004; Goldwater and McClosky, 2005) and compound splitting (Koehn and Knight, 2003; Stymne, 2008; Stymne, 2009) methods. Our approach here is similar, yet uses different implementations; we also studied the joint effect of combining both techniques.

3.1 Reducing the lexical redundancy

In German, determiners, pronouns, nouns and adjectives carry inflection marks (typically suffixes)

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 54–59,
Uppsala, Sweden, 15-16 July 2010.

Input	POS	Lemma	Analysis
In	APPR	in	APPR.In
der*	ART	d	ART.Def.Dat.Sg.Fem
Folge	NN	Folge	N.Reg.Dat.Sg.Fem
befand	VVFIN	befinden	VFIN.Full.3.Sg.Past.Ind
die*	ART	d	ART.Def.Nom.Sg.Fem
derart	ADV	derart	ADV
gestärkte*	ADJA	gestärkt	ADJA.Pos.Nom.Sg.Fem
Justiz	NN	Justiz	N.Reg.Nom.Sg.Fem
wiederholt	ADJD	wiederholt	ADJD.Pos
gegen	APPR	gegen	APPR.Acc
die*	ART	d	ART.Def.Acc.Sg.Fem
Regierung	NN	Regierung	N.Reg.Acc.Sg.Fem
und	KON	und	CONJ.Coord.-2
insbesondere	ADV	insbesondere	ADV
gegen	APPR	gegen	APPR.Acc
deren*	PDAT	d	PRO.Dem.Subst.-3.Gen.Sg.Fem
Geheimdienste*	NN	Geheimdienst	N.Reg.Acc.Pl.Masc
.	$.	.	SYM.Pun.Sent

Table 1: TreeTagger and RFTagger outputs. Starred word forms are modified during preprocessing.

so as to satisfy agreement constraints. Inflections vary according to gender, case, and number information. For instance, the German definite determiner could be marked in sixteen different ways according to the possible combinations of genders (3), case (4) and number (2)[1], which are fused in six different tokens *der, das, die, den, dem, des*. With the exception of the plural and genitive cases, all these words translate to the same English word: *the*. In order to reduce the size of the German vocabulary and to improve the robustness of the alignment probabilities, we considered various normalization strategies for the different word classes. In a nutshell, normalizing amounts to collapsing several German forms of a given lemma into a unique representative, using manually written normalization patterns. A pattern typically specifies which forms of a given morphological paradigm should be considered equivalent when translating into English. These normalization patterns use the lemma information computed by the TreeTagger and the fine-grained POS information computed by the RFTagger (Schmid and Laws, 2008), which uses a tagset containing approximately 800 tags. Table 1 displays the analysis of an example sentence. [2]

In most cases, normalization patterns replace a word form by its lemma; in order to partially preserve some inflection marks, we introduced two generic suffixes, *+s* and *+en* which respectively denote plural and genitive wherever needed. Typical normalization rules take the following form:

- For articles, adjectives, and pronouns (Indefinite , possessive, demonstrative, relative and reflexive), if a token has;
  - Genitive case: replace with lemma+en (Ex. *des*, *der*, *des*, *der* → *d+en*)
  - Plural number: replace with lemma+s (Ex. *die*, *den* → *d+s*)
  - All other gender, case and number: replace with lemma (Ex. *der*, *die*, *das*, *die* → *d*)
- For nouns;
  - Plural number: replace with lemma+s (Ex. *Bilder*, *Bildern*, *Bilder* → *Bild+s*))
  - All other gender and case: replace with lemma (Ex *Bild*, *Bilde*, *Bildes* → *Bild*;

Using these tags, a normalized version of previous sentence is as follows: *In d Folge befand d derart gestärkt Justiz wiederholt gegen d Regierung und insbesondere gegen d+en Geheimdienst+s*. Several experiments were carried out to assess the effect of different normalization schemes. Removing all gender and case information, except for the genitive for articles, adjectives and pronouns, allowed to achieve the best *BLEU* scores.

3.2 Compound Splitting

Combining nouns, verbs and adjectives to forge new words is a very common process in German.

[1] For the plural forms, gender distinctions are neutralized and the same 4 forms are used for all genders .

[2] The English reference: *Subsequently , the energized judiciary continued ruling against government decisions , embarrassing the government – especially its intelligence agencies* .

It partly explains the difference between the number of types and tokens between English and German in parallel texts. In most cases, compounds are formed by a mere concatenation of existing word forms, and can easily be split into simpler units. As words are freely conjoined, the vocabulary size increases vastly, yielding to sparse data problems that turn into unreliable parameter estimates. We used the frequency-based segmentation algorithm initially introduced in (Koehn and Knight, 2003) to handle compounding. Our implementation extends this technique to handle the most common letter fillers at word junctions. In our experiments, we investigated different splitting schemes in a manner similar to the work of (Stymne, 2008).

4 French-English systems

4.1 Baseline N-coder systems

For this language pair, we used our in-house *N-code* system, which implements the *n*-gram-based approach to SMT. In a nutshell, the translation model is implemented as a stochastic finite-state transducer trained using a *n*-gram model of (source,target) pairs (Casacuberta and Vidal, 2004). Training this model requires to reorder source sentences so as to match the target word order. This is performed by a stochastic finite-state reordering model, which uses part-of-speech information[3] to generalize reordering patterns beyond lexical regularities.

In addition to the translation model, our system implements eight feature functions which are optimally combined using a discriminative training framework (Och, 2003): a *target-language model*; two *lexicon models*, which give complementary translation scores for each tuple; two *lexicalized reordering models* aiming at predicting the orientation of the next translation unit; a 'weak' distance-based *distortion model*; and finally a *word-bonus model* and a *tuple-bonus model* which compensate for the system preference for short translations. One novelty this year are the introduction of lexicalized reordering models (Tillmann, 2004). Such models require to estimate reordering probabilities for each phrase pairs, typically distinguishing three case, depending whether the current phrase is translated *monotone*, *swapped* or *discontiguous* with respect to the previous (respectively next phrase pair).

In our implementation, we modified the three orientation types originally introduced and consider: a *consecutive* type, where the original monotone and swap orientations are lumped together, a *forward* type, specifying a discontiguous forward orientation, and a *backward* type, specifying a discontiguous backward orientation. Empirical results showed that in our case, the new orientations slightly outperform the original ones. This may be explained by the fact that the model is applied over tuples instead of phrases.

Counts of these three types are updated for each unit collected during the training process. Given these counts, we can learn probability distributions of the form $p_r(orientation|(st))$ where $orientation \in \{c, f, b\}$ (consecutive, forward and backward) and (st) is a translation unit. Counts are typically smoothed for the estimation of the probability distribution.

The overall search process is performed by our in-house *n-code* decoder. It implements a beam-search strategy on top of a dynamic programming algorithm. Reordering hypotheses are computed in a preprocessing step, making use of reordering rules built from the word reorderings introduced in the tuple extraction process. The resulting reordering hypotheses are passed to the decoder in the form of word lattices (Crego and no, 2006).

4.2 A bilingual POS-based reordering model

For this year evaluation, we also experimented with an additional reordering model, which is estimated as a standard n-gram language model, over *generalized translation units*. In the experiments reported below, we generalized tuples using POS tags, instead of raw word forms. Figure 1 displays the same sequence of tuples when built from surface word forms (top), and from POS tags (bottom).

we	*want*	*translations*	*perfect*
nous	*voulons*	*des_traductions*	*parfaites*

pronoun	*verb*	*noun*	*adjective*
pronoun	*verb*	*det_noun*	*adjective*

Figure 1: *Sequence of units built from surface word forms (top) and POS-tags (bottom).*

Generalizing units greatly reduces the number of symbols in the model and enables to take larger

[3] Part-of-speech information for English and French is computed using the above mentioned TreeTagger.

n-gram contexts into account: in the experiments reported below, we used up to 6-grams. This new model is thus helping to capture the mid-range syntactic reorderings that are observed in the training corpus. This model can also be seen as a translation model of the sentence structure. It models the adequacy of translating sequences of source POS tags into target POS tags. Additional details on these new reordering models can be found in (Crego and Yvon, 2010).

4.3 Combining translation models

Our main translation model being a conventional n-gram model over bilingual units, it can directly take advantage of all the techniques that exist for these models. To take the diversity of the available parallel corpora into account, we independently trained several translation models on subpart of the training data. These translation models were then linearly interpolated, where the interpolation weights are chosen so as to minimize the perplexity on the development set.

5 Language Models

The English and French language models (LMs) are the same as for the last year's French-English task (Allauzen et al., 2009) and are heavily tuned to the newspaper/newswire genre, using the first part of the WMT09 official development data (dev2009a). We used all the authorized news corpora, including the French and English Gigaword corpora, for translating both into French (1.4 billion tokens) and English (3.7 billion tokens). To estimate such LMs, a vocabulary was defined for both languages by including all tokens in the WMT parallel data. This initial vocabulary of 130K words was then extended with the most frequent words observed in the training data, yielding a vocabulary of one million words in both languages. The training data was divided into several sets based on dates and genres (resp. 7 and 9 sets for English and French). On each set, a standard 4-gram LM was estimated from the 1M word vocabulary with in-house tools using Kneser-Ney discounting interpolated with lower order models (Kneser and Ney, 1995; Chen and Goodman, 1998)[4]. The resulting LMs were then linearly combined using interpolation coefficients chosen so as to minimize perplexity of the development set (dev2009a). The final LMs were finally pruned using perplexity as pruning criterion (Stolcke, 1998).

For German, since we have less training data, we only used the German monolingual texts (Europarl-v5, News Commentary and News Monolingual) provided by the organizers to train a single n-gram language model, with modified Kneser-Ney smoothing scheme (Chen and Goodman, 1998), using the SRILM toolkit (Stolcke, 2002).

6 Tuning

Moses-based systems were tuned using the implementation of minimum error rate training (MERT) (Och, 2003) distributed with the Moses decoder, using the development corpus (news-test2008).

The *N-code* systems were also tuned by the same implementation of *MERT*, which was slightly modified to match the requirements of our decoder. The *BLEU* score is used as objective function for MERT and to evaluate test performance. The interpolation experiment for French-English was tuned on news-test2008a (first 1025 lines). Optimization was carried out over newstest2008b (last 1026 lines).

7 Experiments

For each system, we used all the available parallel corpora distributed for this evaluation. We used *Europarl* and *News commentary* corpora for German-English task and *Europarl, News commentary, United Nations* and *Gigaword* corpora for the French-English tasks. All corpora were aligned with *GIZA++* for word-to-word alignments with *grow-diag-final-and* and default settings. For the German-English tasks, we applied normalization and compound splitting as a preprocessing step. For the French-English tasks, we used new POS-based reordering model and interpolation.

7.1 German-English Tasks

We combined our two preprocessing schemes (see Section 3) by applying compound splitting over normalized data. Our experiments showed that for German to English, using 4 characters as the minimum split length and 8 characters as the minimum compound candidate, and allowing the insertion of *-s -n -en -nen -e -es -er -ien)* and the truncation of

[4]Given the amount of training data, the use of the modified Kneser-Ney smoothing is prohibitive while previous experiments did not show significant improvements.

-e -en -n yielded the best *BLEU* scores. On the reverse direction, the best setting is different: 5 characters as minimum split length, 10 characters as minimum compound candidate, no truncation.

These processes are performed before alignment, training, tuning and decoding. Before decoding, we also replaced all OOV words with their lemma. We used the Moses (Koehn et al., 2007) decoder, with default settings, to obtain the translations. For translating from English to German, we used a two-level decoding. The first decoding step translates English to "preprocessed German", which is then turned into German by undoing the effect of normalization. In this second step, we thus aim at restoring inflection marks and at merging compounds. For this second "translation" step, we also use a Moses-based system. To point out the error rate of the second step, we also translated the preprocessed reference German text and computed the *BLEU* score as 97.05. Our experiments showed that this two-level decoding strategy was not improving the direct baseline systems. Table 2 reports the *BLEU* scores[5] on *newstest2010* of our official submissions.

System	*De* $\rightarrow$ *En*	*En* $\rightarrow$ *De*
Baseline	20.0	15.3
Norm+Split	21.3	15.0

Table 2: Results for German-English

7.2 French-English tasks

As explained above, in addition to the baseline system (**base**), two contrast systems were built. The first introduces an additional POS-based bilingual 6-gram reordering model (**bilrm**), the second implements the bilingual n-gram model after interpolating 4 models trained respectively on the news, epps, UNdoc and gigaword subparts of the parallel corpus (**interp**). Optimization was carried out over newstest2008b (last 1026 lines) and tested over newstest2010 (2489 lines). Table 3 reports translation accuracy for the three systems and for both translation directions.

As can be seen, the system using the new reordering model (base+bilrm) outperformed the baseline system when translating into French, while no difference was measured when translating into English. The interpolation experiments did not show any clear impact on performance.

[5]Scores are computed with the official script mteval-v11b.pl

System	*Fr* $\rightarrow$ *En*	*En* $\rightarrow$ *Fr*
base	26.52	27.22
base+bilrm	26.50	27.84
base+bilrm+interp	26.84	27.62

Table 3: Results for French-English

8 Conclusions

In this paper, we presented our statistical MT systems developed for the WMT'10 shared task, including several novelties, namely the preprocessing of German, and the integration of several new techniques in our n-gram based decoder.

Acknowledgments

This work was partly realized as part of the Quaero Program, funded by OSEO, the French agency for innovation.

References

Alexandre Allauzen, Josep M. Crego, Aurélien Max, and François Yvon. 2009. LIMSI's statistical translation systems for WMT'09. In *Proceedings of WMT'09*, Athens, Greece.

Francesco Casacuberta and Enrique Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

Stanley F. Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Simon Corston-oliver and Michael Gamon. 2004. Normalizing german and english inflectional morphology to improve statistical word alignment. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, pages 48–57. Springer Verlag.

Josep M. Crego and José B. Mari no. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Daniel Déchelotte, Gilles Adda, Alexandre Allauzen, Olivier Galibert, Jean-Luc Gauvain, Hélène Meynard, and François Yvon. 2008. LIMSI's statistical translation systems for WMT'08. In *Proc. of the NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of Human Language Technology*

Conference and Conference on Empirical Methods in Natural Language Processing, pages 676–683, Vancouver, British Columbia, Canada, October.

Reinhard Kneser and Herman Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP'95*, pages 181–184, Detroit, MI.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 187–193. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

José B. Mariño, Rafael E. Banchs R, Josep M. Crego, Adrià de Gispert, Patrick Lambert, José A.R. Fonollosa, and Marta R. Costa-Jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Sonja Niessen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntatic information. *Computational Linguistics*, 30(2):181–204.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester, UK, August. Coling 2008 Organizing Committee.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Andreas Stolcke. 1998. Entropy-based pruning of backoff language models. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

Sara Stymne. 2008. German compounds in factored statistical machine translation. In *GoTAL '08: Proceedings of the 6th international conference on Advances in Natural Language Processing*, pages 464–475, Berlin, Heidelberg. Springer-Verlag.

Sara Stymne. 2009. A comparison of merging strategies for translation of german compounds. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 61–69, Morristown, NJ, USA. Association for Computational Linguistics.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics 2004*, pages 101–104, Boston, MA, USA.