**LIMSI’s statistical translation systems for WMT’10**

**Alexandre Allauzen, Josep M. Crego, I˙lknur Durgar El-Kahlout and Franc¸ois Yvon**

LIMSI/CNRS and Universite´ Paris-Sud 11, France BP 133, 91403 Orsay Cedex [Firstname.Lastname@limsi.fr](mailto:Firstname.Lastname@limsi.fr)

**Abstract**

This paper describes our Statistical Ma- chine 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 nor- malizing the German side through a proper preprocessing, aimed at reducing the lex- ical redundancy and at splitting complex compounds. For French-English, we stud- ied 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 exper- iments, we report the improvements ob- tained on the development and test data.

**1 Introduction**

LIMSI took part in the WMT 2010 evalua- tion 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 re- dundancy 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* (Marin˜ o et al., 2006) decoder.

**2 System architecture and resources**

In this section, we describe the main characteris- tics 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 (De´chelotte et al., 2008) which have demonstrated that better nor- malization tools provide better *BLEU* scores (Pap- ineni 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) tok- enizer. 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 treat- ing each word form independently from the oth- ers is plagued with data sparsity, which poses a number of difficulties both at training and de- coding time. When aligning parallel texts at the word level, German compound words typi- cally tend to align with more than one English word; this, in turn, tends to increase the number of possible translation counterparts for each En- glish type, and to make the corresponding align- ment scores less reliable. In decoding, new com- pounds or unseen morphological variants of ex- isting 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; Goldwa- ter 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 ad- jectives carry inflection marks (typically suffixes)

54

*Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 54–59, Uppsala, Sweden, 15-16 July 2010. *Q*c 2010 Association for Computational Linguistics

|  |  |
| --- | --- |
| **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 |
| gesta¨rkte\* | ADJA gesta¨rkt 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 infor- mation. For instance, the German definite deter- miner 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 gen- itive 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 robust- ness 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 manu- ally written normalization patterns. A pattern typ- ically specifies which forms of a given morpho- logical paradigm should be considered equivalent when translating into English. These normaliza- tion patterns use the lemma information computed by the TreeTagger and the fine-grained POS infor- mation computed by the RFTagger (Schmid and Laws, 2008), which uses a tagset containing ap- proximately 800 tags. Table 1 displays the analy- sis of an example sentence. 2

In most cases, normalization patterns replace a word form by its lemma; in order to partially pre-

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 judi- ciary continued ruling against government decisions , embar- rassing the government – especially its intelligence agencies*

.

serve some inflection marks, we introduced two generic suffixes, *+s* and *+en* which respectively denote plural and genitive wherever needed. Typ- ical normalization rules take the following form:

*•* For articles, adjectives, and pronouns (Indef-

inite , 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: re-

place 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 previ- ous sentence is as follows: *In d Folge befand d de- rart gesta¨ rkt 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. Remov- ing all gender and case information, except for the genitive for articles, adjectives and pronouns, al- lowed 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.

It partly explains the difference between the num- ber of types and tokens between English and Ger- man 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 vocab- ulary size increases vastly, yielding to sparse data problems that turn into unreliable parameter esti- mates. We used the frequency-based segmenta- tion algorithm initially introduced in (Koehn and Knight, 2003) to handle compounding. Our im- plementation extends this technique to handle the most common letter fillers at word junctions. In our experiments, we investigated different split- ting 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 transla- tion 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 information3 to generalize reordering patterns be- yond lexical regularities.

In addition to the translation model, our sys- tem implements eight feature functions which are optimally combined using a discriminative train- ing framework (Och, 2003): a *target-language model*; two *lexicon models*, which give comple- mentary translation scores for each tuple; two *lexicalized reordering models* aiming at predict- ing the orientation of the next translation unit;

previous (respectively next phrase pair).

In our implementation, we modified the three orientation types originally introduced and con- sider: a *consecutive* type, where the original monotone and swap orientations are lumped to- gether, a *forward* type, specifying a discontiguous forward orientation, and a *backward* type, specify- ing a discontiguous backward orientation. Empir- ical results showed that in our case, the new orien- tations slightly outperform the original ones. This may be explained by the fact that the model is ap- plied 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 dis-

tributions of the form *pr* (*orientation|*(*st*)) where

*orientation ∈ {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 re- ordering 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 esti- mated 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 sur- face word forms (top), and from POS tags (bot- tom).

a ’weak’ distance-based *distortion model*; and

finally a *word-bonus model* and a *tuple-bonus model* which compensate for the system prefer- ence for short translations. One novelty this year are the introduction of lexicalized reordering mod-



els (Tillmann, 2004). Such models require to



estimate reordering probabilities for each phrase

pairs, typically distinguishing three case, depend- ing whether the current phrase is translated *mono- tone*, *swapped* or *discontiguous* with respect to the

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

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

*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 train- ing corpus. This model can also be seen as a trans- lation 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 perplex- ity 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 Gi- gaword corpora, for translating both into French (1.4 billion tokens) and English (3.7 billion to- kens). To estimate such LMs, a vocabulary was defined for both languages by including all to- kens in the WMT parallel data. This initial vo- cabulary 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

4 Given the amount of training data, the use of the modi- fied Kneser-Ney smoothing is prohibitive while previous ex- periments did not show significant improvements.

chosen so as to minimize perplexity of the de- velopment set (dev2009a). The final LMs were finally pruned using perplexity as pruning crite- rion (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 Good- man, 1998), using the SRILM toolkit (Stolcke,

2002).

**6 Tuning**

Moses-based systems were tuned using the imple- mentation 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 perfor- mance. The interpolation experiment for French- English was tuned on news-test2008a (first 1025 lines). Optimization was carried out over new- stest2008b (last 1026 lines).

**7 Experiments**

For each system, we used all the available par- allel corpora distributed for this evaluation. We used *Europarl* and *News commentary* corpora for German-English task and *Europarl, News com- mentary, United Nations* and *Gigaword* corpora for the French-English tasks. All corpora were aligned with *GIZA++* for word-to-word align- ments with *grow-diag-final-and* and default set- tings. For the German-English tasks, we applied normalization and compound splitting as a pre- processing step. For the French-English tasks, we used new POS-based reordering model and inter- polation.

**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 mini- mum 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

*-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 align- ment, training, tuning and decoding. Before de- coding, we also replaced all OOV words with their lemma. We used the Moses (Koehn et al., 2007) decoder, with default settings, to obtain the trans- lations. 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 merg- ing 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 com- puted 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* scores5 on *newstest2010* of our official submissions.

*System De → En En → 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 bilin- gual 6-gram reordering model (**bilrm**), the second implements the bilingual *n*-gram model after in- terpolating 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 translat- ing into English. The interpolation experiments

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

|  |
| --- |
| *System Fr → En En → 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

did not show any clear impact on performance.

**8 Conclusions**

In this paper, we presented our statistical MT sys- tems developed for the WMT’10 shared task, in- cluding several novelties, namely the preprocess- ing 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, Aure´lien Max, and Franc¸ois Yvon. 2009. LIMSI’s statistical trans- lation systems for WMT’09. In *Proceedings of WMT’09*, Athens, Greece.

Francesco Casacuberta and Enrique Vidal. 2004. Ma- chine 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 lan- guage modeling. Technical Report TR-10-98, Com- puter Science Group, Harvard University.

Simon Corston-oliver and Michael Gamon. 2004.

Normalizing german and english inflectional mor- phology 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 Jose´ B. Mari no. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Daniel De´chelotte, Gilles Adda, Alexandre Allauzen, Olivier Galibert, Jean-Luc Gauvain, He´le`ne Mey- nard, and Franc¸ois Yvon. 2008. LIMSI’s statisti- cal translation systems for WMT’08. In *Proc. of the NAACL-HTL Statistical Machine Translation Work- shop*, Columbus, Ohio.

Sharon Goldwater and David McClosky. 2005. Im- proving statistical MT through morphological analy- sis. 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 *Pro- ceedings of the International Conference on Acous- tics, 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: Pro- ceedings of the tenth conference on European chap- ter of the Association for Computational Linguistics*, pages 187–193. Association for Computational Lin- guistics.

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 Compu- tational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Jose´ B. Marin˜ o, Rafael E. Banchs R, Josep M. Crego, Adria` de Gispert, Patrick Lambert, Jose´ A.R. Fonol- losa, and Marta R. Costa-Jussa`. 2006. N-gram- based machine translation. *Computational Linguis- tics*, 32(4):527–549.

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

Franz J. Och. 2003. Minimum error rate train- ing in statistical machine translation. In *Proceed- ings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sap- poro, Japan.

Helmut Schmid and Florian Laws. 2008. Estima- tion 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 Or-

ganizing Committee.

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

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

Andreas Stolcke. 2002. SRILM – an extensible lan- guage modeling toolkit. In *Proceedings of the Inter- national 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: Pro- ceedings of the 6th international conference on Ad- vances in Natural Language Processing*, pages 464–

475, Berlin, Heidelberg. Springer-Verlag.

Sara Stymne. 2009. A comparison of merging strate- gies for translation of german compounds. In *EACL*

*’09: Proceedings of the 12th Conference of the*

*European Chapter of the Association for Compu- tational 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 *Pro- ceedings of the Human Language Technology con- ference / North American chapter of the Association for Computational Linguistics 2004*, pages 101–104, Boston, MA, USA.