The INESC-ID IWSLT07 SMT System

João V. Graça , Diamantino Caseiro, Luisa Coheur

Spoken Language Lab (L^2F)
Inesc-ID, Lisboa

`javg,dcaseiro,luisa.coheur@l2f.inesc-id.pt`

Abstract

We present the machine translation system used by L^2F from INESC-ID in the evaluation campaign of the International Workshop on Spoken Language Translation (2007), in the task of translating spontaneous conversations in the travel domain from Italian to English.

1. Introduction

This paper describes the machine translation system used by INESC-ID in its first participation on the evaluation campaign of the International Workshop on Spoken Language Translation 2007.

We submitted translation results for manual and first-best transcriptions in the Italian-to-English language pair.

The statistical machine translation system consists of a first-pass phrase-based system using moses machine translation toolkit [1], followed by a reranking step.

In section 2 we describe the system as well as the corpora and the baseline results; in section 3 we present several experiments we did in order to improve the results. Then, in section 4 we show the results we obtained. Finally, section 5 concludes and discusses future work.

2. Overall System Description

2.1. Architecture

The INESC-ID IWSLT07 Statistical Machine Translation (SMT) System architecture is shown in Figure 1. It consists of a pipeline with the following steps: preprocessing, phrase-based first pass decoding, n-best reranker and post-processing.

The first pass module follows the baseline suggested for the ACL second workshop on machine translation[1]. The used features include both direct and inverse phrase probability, IBM1 model lexica over all possible alignments, and phrase and word penalties. Features are combined using a log linear model optimized to maximize BLEU [2]. In this paper, we focus our description on the remaining modules.

[1] http://www.statmt.org/wmt07/

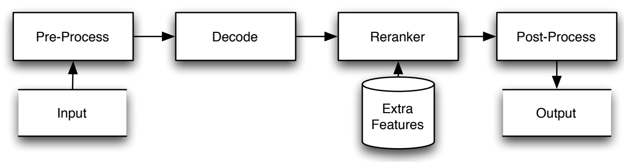

Figure 1: System

2.2. Corpora

Tables 1 to 3 provide corpora partition and description. One of the characteristics of IWSLT evaluations is the reduced size of the training corpus when compared to other MT evaluations.

	Italian	English
Sentences	19845	
Words	14365	184134
Tokens	10126	7011
Avg. Sentence Len.	7.23	9.27

Table 1: Training Corpus

From Table 2 to 3 we can see that the average percentage of unknown words is around 10% which represents an additional burden to the already hard task of translation. Also this years task has additional difficulties. First, the source language (in our case Italian) is lowercase and without punctuation; then the corpus is not separated one sentence per line, and there are lines with different number of sentences that must be translated to one another; finally the training corpus is not speech transcription, so we have a kind of domain adaptation between the training corpus and the test corpus.

3. Experiments

3.1. Baseline results

In order to investigate the difficulty of this year task, a singlepass baseline system was trained on the training set and tested on the various development sets. Table 4 shows the results in each set. We can see that this year's task of spontaneous speech translation is by far the most challenging one.

	Italian	English
Dev1 (IWSLT05 Written)		
Sentences	506*7	
Words	2906	3769
Tokens	1005	929
Avg. Sentence Len.	5.74	7.44
Out of Vocab Words	506	506
Out of Vocab Tokens	497	500
Dev2 (IWSLT06 READ)		
Sentences	489	
Words	4976	6391
Tokens	1234	1135
Avg. Sentence Len.	10.1	13.0
Out of Vocab Words	489	489
Out of Vocab Tokens	486	487
Dev3 (IWSLT07 Speech))		
Sentences	996	
Words	8666	9482
Tokens	871	1437
Avg. Sentence Len.	8.7	9.5
Out of Vocab Words	996	996
Out of Vocab Tokens	807	827

Table 2: Development Corpus

	Italian
IE Clean	
Sentences	724
Words	6540
Tokens	735
Avg. Sentence Len.	9.0
Out of Vocab Words	724
Out of Vocab Tokens	613
IE ASR	
Sentences	724
Words	6384
Tokens	726
Avg. Sentence Len.	8.8
Out of Vocab Words	724
Out of Vocab Tokens	618

Table 3: Test Corpus

All experiments described in this paper, from now on, were evaluated on the dev3 set.

	Dev1	Dev2	Dev3
Baseline	56.60	37.19	16.78

Table 4: Degradation of bleu on diferent corpus types

3.2. Corpora addition

In order to mitigate the sparse data problem we collected more data in the travel domain, namely, a dictionary of verb forms and a tourist domain dictionary.

The motivation for using a verb list is the fact that Italian, being a Romance language, is highly inflected, so a significant quantity of verb forms are not available at training time and appear at testing time. To build the dictionary, we started by selecting the infinitive form of every verb present in the training data. Then an online verb conjugator[2] was used to generate most inflected forms. These forms were then translated to English using an off-the-shelf version of Systran and manually verified.

A dictionary of tourism terms was also collected from phrasebooks, the goal of this dictionary was to decrease the number of unknown nouns existing in the development corpus.

Following results in domain adaptation from [3] we tried to incorporate the new data in different ways:

- Language Model: data was added to the language model training;
- Phrase: data was added to the corpus and used in the alignments and the phrase extraction. It was not used in the language model;
- Phrase and Language Model: data was added both on the phrase extraction and on the language model.

System	BLEU
Baseline	16.78
+data LM	16.82
+data Phrase	16.10
+data Phrase and LM	16.98

Table 5: Tests using extra data.

As Table 5 shows, best results were obtained by adding the data to both translation and language models. However and against our expectation the differences are not significant. The added data consists in pairs of verbs/terms in both languages appended to the original corpus, so these new sentences (composed of one word or small compound terms) will not influence the system much. In the Language Model case we only have 1-gram counts for those terms, and as for the phrase base decoder we only have small phrases which will not be preferred by the system. We will have to investigate how to use this source of information more efficiently.

3.3. Pre-Processing

Additionally to the usual procedure of the pre-processing step, we add the following:

[2] http://www.verbix.com

- Abbreviation expansion: the most commonly used abbreviations (such as *Ms.*, *Ms.* and *Sig.*), were expanded for Italian, since they never appear in the speech transcription.

- Tokenization script changes: the tokenization script that ships with moses was adapted to Italian. The changes included the addition of a list of Italian abbreviations and joining the apostrophe to the left (as opposed to English that joins to the right. For instance in English *don't* become *don 't*, meaning *do not*, while in Italian *dov'è* becomes *dov' è* meaning *dove è*). We also add some exceptions, like *o'clock* on English that is a single token. Table 6 show the baseline difference in Bleu using the different tokenization. Using our own tokenization we gain 0.5 bleu points.

- Ponctuation remover: only ponctuation from the italian corpus was removed. We opted to leave it on the English corpus, in order to let the translation system learn how to introduce it.

System	BLUE
Baseline	16.78
New Tokenization	17.22

Table 6: Baseline results using the original tokenization script and our tokenization script

3.4. Phrase based first pass decoding

In the first pass system we performed several experiments to establish the best configuration.

First, the baseline performance was tested varying the language model order as shown in Table 7. All language models were interpolated Kneser-Ney smoothing [4] models estimated using the SRI Language Model Toolkit [5]. The best result was obtained using 5-grams.

Configuration	BLEU
3-gram	16.59
4-gram	16.60
5-gram	16.98
6-gram	16.68

Table 7: Different language models sizes

Moses factored models allow the use of additional information regarding each word. We explored the use of morphological information. The TreeTagger[3] from the Institute for Computational Linguistics of the University of Stuttgart was used to annotate text with part-of-speech (POS) and lemma information. The Italian parameter file provided by Achim Stein was the one we used in our experiments.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Then, several experiments were made using this information (Table 8 contains the results). These results use the original tokenization as well as the extra data.

- Using lemmas for alignment: As described in [6], in order to improve the quality of phrases extracted from word alignments, these can be performed using each word lemma. The idea is to reduce the data sparseness especially on highly inflected languages. In this step we used the Moses factored models to produce the alignments using the lemmas, and then use the corresponding word surfaces on the decoding process.

- Using lemmas for alignment with original training corpus: The dictionary could be affecting the lemma by having too many entries for the same lemma (all verb forms) and overfitting for other possible translation of the lemma.

- Using a Part-Of-Speech Distortion Model: The intuition behind this model is that using the POS tag can be useful to predict the reordering of sentences, and that those statistics would be less sparse and more informative than only based in words.

- Using several Language Models: The intuition in this experience was to add a POS language model in the model to penalize sentences with uncommon POS sequences.

Although using the lemma-based alignments produces better results on tests performed on dev1, they perform worse than the simple baseline on dev3. It is not obvious to us why this happen and some future analysis is required. Also models that use POS reordering and POS the language model perform much worse. This might be due of a weak tagger accuracy, but mostly because spontaneous speech tends not to have a regular POS sequence. These two models also performed worse than the baseline and lemma on dev1.

Configuration	BLUE
Baseline (extra corpus + original tokenization)	16.98
Using lemma for alignment	16.41
Using lemma for alignment plus original corpus	16.79
Using Part-Of-Speech Distortion Model	12.24
Using different Language Models	15.54

Table 8: Different Moses configurations

3.5. Filtered Phrase Table

By analyzing the phrase table from the previous models we noticed a significant number of sentences in English containing a period in the middle, which lead to excessive meaningless punctuation. Accordingly, the phrase table was filtered by removing all phrases with periods or question marks in the middle. Table 9 shows the results obtained for different models by using the new phrase table.

System	Normal	Filtered
Baseline	17.22	17.45
Baseline with extra data	17.32	17.25
Lemma	16.72	16.89
Lemma with extra data	17.30	17.34

Table 9: Phrase table punctuation filtering

It should be noticed that this procedure does not always produce the best results, but we get the best results for the simple baseline by using it.

3.6. Reranker

In this section we present the second pass reranking system that rescores lists of 1000-best hypotheses generated by the best system described in the previous section. We used the following features in the optimization:

- Ratio between target and source sentence length [7]
- Question features [7]
- 3,4,5-grams target word LMs
- 3,4,5-grams target POS LMs
- Direct and Inverse IBM Model 1 lexica
- Part-of-Speech similarity (3.6.1)

These features were combined with the first pass score according to a log-linear model. Combination weights were trained to maximize BLEU on dev1 using the downhill simplex search algorithm [8].

System	BLEU
Baseline	17.45
+ length ration	17.45
+ question features	17. 51
+ word n-gram LMs	17.45
+ POS n-gram LMs	17.38
+IBM1 Dictionary	17.45
+POS similarity	17.57
All Features	17.66

Table 10: Contribution of each feature in rescoring

Table 10 shows the contribution of each feature when optimized in isolation with the first pass system score. We observe that many features are not useful in isolation, however, when used in combination with other features they bring improvements. The small improvements obtained are disappointing and not in line with improvements reported by other researchers [9, 10].

3.6.1. *Part-Of-Speech Similarity Features*

Two novel features, f_1 and f_2, were introduced which provided the best single rescoring improvement.

f_1 relies on computing similarities between POS tags, and assumes that the number of certain morpho-syntactic entities (such as nouns) should be stable in both a sentence and its translation. Accordingly, for each sentence pair, several tags are counted in both sides. Feature f_1 is calculated with the Formula 1, where it_i stands for the count of tag number i for Italian, and en_i the count of the corresponding tag for English.

$$f_1 = \sqrt[2]{\sum_{n=1}^{\#pos} (it_n - en_n)^2} \qquad (1)$$

It should be noted that one single tag in Italian could correspond to several tags in English and vice-versa (for instance NOM in Italian can be either NNS or NN in English), as such, various equivalence classes were defined between Italian and English tags, as shown in Table 11:

Italian	English
NOM	NNS NN
PRO NON	NP NN NNS
CON	CC

Table 11: Italian and English tag equivalence classes

f_2 relies on computing penalty patterns 12, and assumes that certain sequences of tags (patterns) are very unlikely (such as DT DT, where DT stands for determiner). In order to calculate f_2, the unlikely patterns for English from Table 12 were considered.

English Pattern
DT DT
VV VV
IN IN
DT NN JJ

Table 12: English penalty patterns

Then we searched for these patterns in each sentence. Everytime a sentence matched such a pattern a penalty was added. f_2 is the sum of such penalties. This was just a preliminary experiment with this feature, which we still wish to study more and for different languages, since this is a way of introducing some linguistic constraints on the resulting output.

3.7. Post-Processing

The Post-Processing step is responsible for converting the output into the input original format. First, the recaser tool

	simple	post-processed
Baseline	17.45	21.53
Reranked	17.55	21.58

Table 13: Post-processing results

that ships with Moses was used to get a true cased version of the output. Secondly, the output was converted to the original tokenization of the corpus. Finally, some procedural changes were applied to the output in order to correct some common mistakes. These changes were the following:

- Remove leading and trailing commas;
- Add question marks to lines which started with a question word such as *where*;
- Remove wrong question marks from sentences;
- Add period to lines that ended with no punctuation;
- Change good-bye for goodbye;
- Place a period before capitalized special expression, such as *Good morning*;
- Remove extra spaces.

These kinds of changes require a detailed output analysis and sometimes are too specific, but on the whole, they lead to a significant increase of the BLEU score. Table 13 shows the difference in BLEU with this changes. The 4 points increase was larger than any model or reranking variation.

4. Test Set Results

Two translations of the clean and ASR test sets were submitted for evaluation. The primary one was obtained using our best system that consists of preprocessing, 1st pass, reranker and post-processing. The secondary one is similar but without the reranker. The official scores are presented in Table 14.

Condition	Primary system	Secondary system
IE clean	26.57	26.35
IE clean	24.16	24.35

Table 14: Official Test Set Scores

We noticed that the difference between the reranked and the non-reranked versions is larger than observed in the development set, however, the improvement on clean data is still minimal and a degradation is observed in the ASR condition, most likely because the reranking feature weights were optimized in a clean corpus.

5. Conclusions and Future Work

This work presents the INESC-ID SMT system for the IWSLT 2007 evaluation. The system is a phrase-based multipass system based on log-linear combination of multiple features. The results obtained are promising, however some modules still need improving. One such module is the feature weights optimizer using in rescoring. The downhill simplex algorithm used is very sensitive to the starting point and has difficulty optimizing large number of weights. In the future, we plan to investigate the use of other algorithms and establish an effective strategy for gradually adding features. We also want to understand why some approaches that have showed promising results in other works have not produce such results in our case.

6. Acknowledgements

This work was partially funded by FCT project WFST (POSI/PLP/47175/2002) and INESC-ID Lisboa has support from the POSI Program of the "Quadro Comunitrio de Apoio III.

7. References

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180.

[2] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, E. Hinrichs and D. Roth, Eds., 2003, pp. 160–167.

[3] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 224–227.

[4] R.Kneser and H.Ney, "Improved backing-off for m-gram language modeling," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, Detroit, MI, May 1995, pp. 181–184.

[5] A. Stolcke, "Srilm – an extensible language modeling toolkit," in *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, Denver, 2002, pp. 901–904.

[6] M. Popović and H. Ney, "Improving word alignment quality using morpho-syntactic information," in *COL-*

ING '04: Proceedings of the 20th international conference on Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 2004, p. 310.

[7] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico, "The itc-irst smt system for iwslt-2006," in *International Workshop on Spoken Language Translation Evaluation Campaign on Spoken Language Translation:IWSLT-2006*, Kyoto, Japan, November 2006, pp. 53–58.

[8] J. Nelder and R. Mead, "A simplex method for function minimization," *Computing Journal*, vol. 4, no. 7, pp. 308–313, 1965.

[9] L. Shen, A. Sarkar, and F. J. Och, "Discriminative reranking for machine translation," in *HLT-NAACL*, 2004, pp. 177–184.

[10] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. R. Radev, "A smorgasbord of features for statistical machine translation," in *HLT-NAACL*, 2004, pp. 161–168.