An Italian to Catalan RBMT system reusing data from existing language pairs*

Antonio Toral
School of Computing
Dublin City University
Ireland
atoral@computing.dcu.ie

Mireia Ginestí-Rosell
Prompsit Language Engineering
E-03195 l'Altet (Spain)
mireia@prompsit.com

Francis Tyers
DLSI
Universitat d'Alacant
E-03070 Alacant
ftyers@dlsi.ua.es

Abstract

This paper presents an Italian→Catalan RBMT system automatically built by combining the linguistic data of the existing pairs Spanish–Catalan and Spanish–Italian. A lightweight manual postprocessing is carried out in order to fix inconsistencies in the automatically derived dictionaries and to add very frequent words that are missing according to a corpus analysis. The system is evaluated on the KDE4 corpus and outperforms Google Translate by approximately ten absolute points in terms of both TER and GTM.

1 Introduction

One of the most common criticisms towards Rule-Based Machine Translation (RBMT) regards the amount of work necessary to build a system for a new language pair (Somers, 2003). In fact, in a traditional scenario, linguists with expertise in the source and target language need to manually build all the dictionary entries and transfer rules. Conversely, in the Statistical Machine Translation (SMT) approach (Koehn, 2010), no such effort is required as the system can be automatically built from parallel corpora. However, this approach is only applicable for those language pairs for which big amounts of parallel text are available.

In this paper we present an automatically built RBMT system by exploiting linguistic data from existing language pairs. Our approach builds an MT system for a language pair a–b given existing systems for the language pairs a–c and b–c. Specifically, we have built a new language pair for the Apertium RBMT engine, Italian–Catalan, by exploiting the existing Spanish–Italian and Catalan–Spanish language pairs. It is worth mentioning the lack of parallel resources for Catalan (e.g. Europarl (Koehn, 2005) is the most widely used resource of parallel documents for European languages, but it does not cover Catalan).

Our motivation can be then summarised by the following two basic ideas:

- RBMT is a competitive and useful approach for those languages for which there are no parallel corpora available (Forcada, 2006).
- Reutilising data from similar existing language pairs can significantly reduce the amount of work required to build a new language pair.

The rest of the paper is structured as follows. The following section presents the RBMT system Apertium, emphasising on approaches that consider reuse of resources and automatic acquisition of linguistic data. After that we introduce our methodology. Subsequently, we provide the evaluation of the created system, and compare its performance to a state-of-the-art SMT engine. Finally we outline some conclusions and propose lines of future work.

2 Background

Apertium is an open-source rule-based machine translation platform initially built for related lan-

*This research has been partially funded by the EU project PANACEA (7FP-ITC-248064).

F. Sánchez-Martínez, J.A. Pérez-Ortiz (eds.)
Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 77–81
Barcelona, Spain, January 2011. http://hdl.handle.net/10609/5650

guage pairs (such as Spanish–Portuguese), but later expanded to deal with more divergent pairs. It uses finite-state transducers (Roche and Schabes, 1997) for lexical processing, hidden Markov models for part-of-speech tagging (Cutting et al., 1992), and multi-stage finite-state *chunking* for structural transfer.

The linguistic data needed to create a machine translation system between two languages in Apertium are: morphological dictionaries for the source language and for the target language, a bilingual dictionary, structural transfer rules and a tagger definition file with optional linguistic restrictions to train an optimal statistical part-of-speech tagger.

Since its first version in 2005, the number of language pairs available has grown steadily and today (as of 20th November, 2010) there are 25 released stable language pairs, with stable linguistic resources for 20 languages[1]; there are also preliminary linguistic resources for some more languages (including Italian). A large community has grown around it and there are contributors in many different countries.

Linguistic resources are encoded in standard formats, which eases its reuse for new translation pairs and for other language technologies. Source-language morphological dictionaries are theoretically independent from the target language, although in practice some bias does exist towards the target language; bilingual dictionaries and structural transfer rules have to be created specifically for each translation pair.

Several papers describe the creation of data for new Apertium language pairs, using a variety of approaches, including the reuse of existing free/open source resources (Sánchez-Martínez et al., 2008; Sánchez-Martínez and Forcada, 2009; Ginestí-Rosell et al., 2009; Tyers et al., 2009; Tyers and Donnelly, 2009; Unhammer and Trosterud, 2009) and the use of Crossdics (Armentano and Forcada, 2008),[2] a program provided in the Apertium platform that, given two existing systems between the language pairs *a*–*c* and *b*–*c*, is used to obtain dictionaries for a new translation pair *a*–*b*. This is the method we used to create the Italian–Catalan translation pair, using the available Apertium translation pairs Spanish–Italian and Spanish–Catalan. According to (Armentano and Forcada, 2008), using Crossdics to cross dictionaries and adding some manual work to correct and improve the resulting data is a good and fast starting point for a new translation pair.

3 Methodology

The overall process is depicted in figure 1. First, Crossdics is applied to the Spanish–Italian and Spanish–Catalan language pairs to automatically derive linguistic data for Italian–Catalan. Subsequently, the created dictionaries are automatically analysed to detect inconsistencies, which are manually corrected. Finally, we extract the most frequent words from Italian corpora and add them to our dictionaries in order to improve the coverage of the translation system. The next paragraphs present in more detail each of the phases.

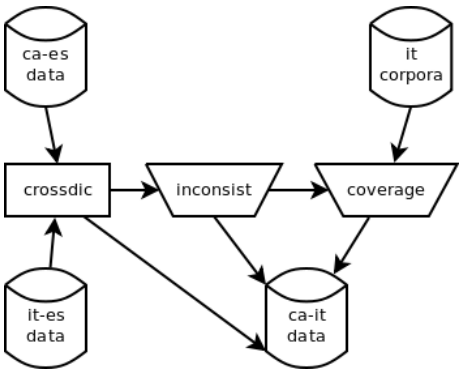

Figure 1: Method diagram

Our starting point was the Spanish–Italian (`es-it`) and Spanish–Catalan (`es-ca`) language pairs in Apertium. The `es-it` pair is not a stable one and therefore has not yet been released; the `es-ca` pair is stable and its last released version is number 1.2.0 from October 2009. For both language pairs, we used the latest revision from the svn in Sourceforge to have the most up-to-date data: revision number 25524 (15th September, 2010) for `es-it` and revision number 26005 (8th October, 2010) for `es-ca`.

As for the first pair, the Italian monolingual dictionary contains 10,351 entries,[3] the Spanish

[1] `wiki.apertium.org`

[2] `http://wiki.apertium.org/wiki/Crossdics`

[3] an entry in the Apertium dictionaries consists of a lemma and its inflection paradigm

monolingual dictionary 11,419 entries, and the bilingual dictionary, 12,445 correspondences. In the Spanish–Catalan pair, the Spanish monolingual dictionary contains 43,575 entries, the Catalan monolingual dictionary, 40,012 entries, and the bilingual dictionary, 50,735 bilingual correspondences.

We applied Crossdics to these two pairs and automatically obtained three dictionaries: an Italian monolingual dictionary, with 7,415 entries; a Catalan monolingual dictionary, with 8,295 entries; and a Italian–Catalan bilingual dictionary, with 8,726 correspondences. These preliminary dictionaries contained many inconsistencies, mainly due to differences of gender and number between both languages and to different ways of categorising lemmas and morphological features in the two source pairs. We decided that the best way to solve the inconsistencies in the Catalan monolingual dictionary was to substitute this with the Catalan dictionary from the `es-ca` pair, since the latter was more consistent and we found that, using it, less amount of work was needed to correct the errors. This dictionary contained many more terms, around 40,000, but since we intended to build only a translation engine in the Italian–Catalan direction, this fact did not suppose any problem.

The automatically detected dictionary inconsistences were manually solved, and the amount of time needed to complete this task was two weeks by one person.[4]

We calculated then the coverage of the system on two Italian corpus, the Italian Europarl (Koehn, 2005) corpus and the Italian Wikipedia[5]. Table 1 shows the coverage values for these two corpus.

The next step was to add the most frequent unknown words from both corpus. We added a total of 155 entries to the Italian monolingual dictionary, and the necessary bilingual entries in the bilingual dictionary; there was no need to add entries to the Catalan dictionary since it came from the `es-ca` pair and had a very high coverage. The result of this improvement was an increase of 2.5 and 3.9 points in the coverage percentage for the Europarl corpus and the Wikipedia corpus respectively. The figures are shown in table 1.

Once the dictionaries were corrected and improved, we added to the system the other required linguistic data files. The tagger definition file and the disambiguation probabilities for Italian were taken directly from the Spanish–Italian pair, with no modifications. The transfer rules were taken from another pair of romanic languages, namely the Occitan–Catalan pair. We took almost all the rules for noun phrases (which are basically responsible for number and gender concordance operations) and some other rules for other word patterns. After this, we created 9 rules manually to deal with some verb constructions and combinations of verbs with clitic pronouns. The number of transfer rules in the final version is: 42 rules in the transfer first submodule file (file `apertium-ca-it.it-ca.t1x`) and 2 rules in the transfer second submodule file (`apertium-ca-it.it-ca.t2x`). No rules for the third submodule were created.

4 Evaluation

This section presents the evaluation. First, in 4.1 we describe the experimental environment. Then, in 4.2, we show the results obtained and draw conclusions from them.

4.1 Environment

The dataset used for the experiment has been extracted from the KDE4 multilingual documentation corpus in the OPUS project (Tiedemann and Nygard, 2004).[6] Its Italian–Catalan bilingual corpus contains 146,372 sentence pairs. We discarded those where the source or target is shorter than 10 words or longer than 30, those where the difference of number of words is higher than 10% and those that contain URLs, Copyright notices and source code. This leads to a candidate test set of 6,927 sentences, from it we randomly selected 1,000 sentences.

Several state-of-the-art automatic MT metrics are used to assess the performance of each system. Specifically, we use the following ones: TER (Snover et al., 2006) , GTM (Turian et al., 2003), BLEU (Papineni et al., 2002) and

[4] considering five days of work a week, and eight hours a day

[5] `it.wikipedia.org`

[6] `http://urd.let.rug.nl/tiedeman/OPUS/KDE4v2.php`

	Italian Europarl	Italian Wikipedida
Number of tokenised words	46,569,602	241,563,615
Coverage before adding most frequent words	86.4%	75.5%
Coverage after adding most frequent words	88.9%	79.4%

Table 1: Coverage of Italian monolingual dictionary on two corpus

NIST (Doddington, 2002). Statistical significance tests are carried out for BLEU and NIST (with ARK's code)[7] and for GTM (using FastMtEval).[8] P-value is set to 0.01.

The following systems are evaluated:

- Apertium, is the Italian→Catalan translator developed in this paper.
- Apertium-i, performs the translation indirectly using the already existing Apertium engines Italian→Spanish and Spanish→Catalan.
- Google Translate,[9] is a state-of-the-art general-purpose on-line Statistical MT system which provides Italian to Catalan translation.

4.2 Results

Table 2 shows the results obtained for the aforementioned experimental setting.

Metric	Apertium	Apertium-i	Google
TER	**0.5703**	0.6118	0.6785
GTM	**0.5162**	0.4712	0.41637
BLEU	0.2290	0.1492	**0.2459**
NIST	5.6567	4.4753	**6.1071**

Table 2: Results

Two different trends can be noticed from these results according to the different metrics. On the one hand, Apertium is approximately ten absolute points over Google for TER and GTM. For these metrics Apertium-i is between the other two systems, roughly four points below Apertium and six over Google. The differences in GTM both between Apertium and Apertium-i and between Apertium-i and Google are significant.

On the other hand, Google is the best system according to BLEU and NIST scores. It is worth mentioning that these metrics are known to be biased towards SMT systems (Callison-Burch and Osborne, 2006). Google obtains 1.69 absolute BLEU points over Apertium but the difference is not statistically significant. With respect to NIST, the difference is of 0.44 points and it is significant. Apertium is significantly better than Apertium-i both for BLEU (7.98 points) and for NIST (1.18 points).

5 Conclusions

This paper has presented an Italian→Catalan RBMT system obtained by automatically deriving its linguistic data from existing Italian–Spanish and Catalan–Spanish systems. Only a limited amount of manual work was carried out to (i) correct the inconsistencies found in the resulting dictionaries, (ii) augment the coverage by adding a limited amount of very frequent lemmas appearing in two Italian corpora (Wikipedia and Europarl) and to (iii) add some transfer rules for general word patterns. The system has been evaluated and its performance compared to (i) indirect translation using the two RBMT engines sequentially (Italian→Spanish→Catalan) and to (ii) a state-of-the-art SMT system. The system presented yields significant improvement over indirect RBMT across all the automatic MT metrics considered. Compared to the SMT system, it obtains significant better scores for the TER and GTM metrics (around 10 absolute points) while obtains comparable performance for NIST and slightly worst for BLEU (1.69 absolute points).

Another contribution of the paper is the availability of the software and data developed. These comprise the Apertium Italian–Catalan linguistic data, software to extract a testset from the KDE OPUS corpus, the testset itself and the system runs. All of the above is available under the GNU General Pub-

[7] http://www.ark.cs.cmu.edu/MT/

[8] http://www.computing.dcu.ie/~nstroppa/index.php?page=softwares

[9] http://translate.google.com

lic License from `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-ca-it`.

References

Armentano, C. and Forcada, M. L. (2008). Reutilización de datos lingüísticos para la creación de un sistema de traducción automática para un nuevo par de lenguas. *Procesamiento del Lenguaje Natural*, 41:243–250.

Callison-Burch, C. and Osborne, M. (2006). Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL*, volume 2006, pages 249–256.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In *3rd Conf. on Applied NLP. Association for Comp. Ling. Proc. of the Conference*, pages 133–140.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Forcada, M. L. (2006). Open-source machine translation: an opportunity for minor languages. In *Proceedings of the 5th SALTMIL workshop on Minority Languages*.

Ginestí-Rosell, M., Ramírez-Sánchez, G., Ortiz-Rojas, S., Tyers, F. M., and Forcada, M. L. (2009). Development of a free Basque to Spanish machine translation system. *Procesamiento del Lenguaje Natural*, (43):187–195.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Roche, E. and Schabes, Y. (1997). Introduction. In Roche, E. and Schabes, Y., editors, *Finite-State Language Processing*, pages 1–65. MIT Press, Cambridge, Mass.

Sánchez-Martínez, F. and Forcada, M. L. (2009). Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34:605–635.

Sánchez-Martínez, F., Pérez-Ortiz, J. A., and Forcada, M. L. (2008). Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, 22(1-2):29–66. DOI: 10.1007/s10590-008-9044-3.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Somers, H., editor (2003). *Computers and Translation: A translator's guide*, chapter Why translation is difficult for computers (by D. Arnold). Benjamins Translation Library.

Tiedemann, J. and Nygard, L. (2004). The OPUS corpus - parallel and free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*, Lisbon, Portugal.

Turian, J., Shen, L., and Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. In *In Proceedings of MT Summit IX*, pages 386–393.

Tyers, F. M. and Donnelly, K. (2009). apertium-cy–a collaboratively-developed free RBMT system for Welsh to English. *Prague Bull. of Math. Ling.*, 91:57–66.

Tyers, F. M., Wiechetek, L., and Trosterud, T. (2009). Developing prototypes for machine translation between two Sámi languages. In *Proc. of the 13th Annual Conf. of the EAMT, EAMT09*, pages 120–128.

Unhammer, K. and Trosterud, T. (2009). Reuse of free resources in machine translation between nynorsk and bokmål. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42, Alicante.