The PRESEMT Project

Adam Kilgarriff, George Tambouratzis
Lexical Computing Ltd., UK; ILSP, Athens, Greece
E-mail: adam@lexmasterclass.com, giorg_t@ilsp.gr

Abstract

Within the PRESEMT project, we have explored a hybrid approach to machine translation in which a small parallel corpus is used to learn mapping rules between grammatical constructions in the two languages, and large target-language corpora are used for refining translations. We have also taken forward methods for 'corpus measurement', including an implemented framework for measuring the distance between any two corpora of the same language. We briefly describe developments in both these areas.

Keywords: hybrid machine translation, machine learning, corpus distance measures, comparing corpora

1. Introduction

PRESEMT (Pattern-Recognition-based Statistically Enhanced Machine Translation) is an EU FP7 Project running from January 2010 to December 2012. It is developing a language-independent methodology for the creation of a flexible and adaptable system which can be ported to new language pairs and specific user requirements with relative ease. Unlike most statistical system, it does not assume that large parallel corpora are available for a given language pair, as they often are not. It uses a small parallel corpus to learn automatically how the syntactic constructions of the source language map to those of the target language, a bilingual dictionary for lexical transfer, and a large monolingual corpus for target-language modelling. As of April 2012, a prototype system is available on the web for the directed language pairs English to German, German to English, and Czech, Greek and Norwegian to German and to English. In the final year of the project the Consortium will port the methodology to new language pairs, involving translating from any of the aforementioned languages to Italian.

Language technology based on machine-learning from corpora will always depend on the nature and quality of the corpus or corpora used for training. With this in mind, the project has also undertaken foundational work in this area. In this extended abstract, we first outline the system and then briefly describe the work performed within the project on corpus comparison.

2. The PRESEMT MT system

This article focuses on the PRESEMT project (www.presemt.eu), which aims to develop a language-independent methodology for creating MT systems. This method overcomes well-known problems of other MT approaches, such as bilingual corpora compilation or creation of new language-specific rules.

Most recent MT approaches adopt the Statistical Machine translation paradigm (Koehn, 2010), where a statistical model is extracted probabilistically from a large parallel corpus to represent the transition from source (SL) to target language (TL). In Statistical Machine Translation, an important bottleneck is the need for extensive bilingual corpora between SL and TL. Though such corpora may exist between widely-used languages, they rarely exist for less widely-used languages, while their construction would require substantial resources.

PRESEMT builds on experience accumulated within the METIS (Dologlou et al., 2003) and METIS-2 (Markantonatou et al., 2006), projects, where the theme was the implementation of MT using solely data from TL monolingual corpora via pattern recognition techniques. Analysing the behaviour of METIS-2, a potential improvement in translation quality was identified. This involved supplementing the monolingual TL corpus with a small bilingual corpus (of typically a few hundred sentences), to provide the basis for the translation output.

The PRESEMT translation process is based on phrases, as that improves the translation quality. Translation is split into two phases, each of which focuses on processing a single type of corpus to resolve specific types of information in the output sentence. Phase 1 (Structure selection) utilises the small bilingual corpus to determine the appropriate TL phrasal structure for input sentences, establishing the order and type of TL phrases. The structure selection output is a sequence of TL structures that contain phrase and tag information and sets of TL lemmas as retrieved from the bilingual dictionary.

Phase 2 (Translation Equivalent Selection) accesses the monolingual corpus to specify the word order within each phrase and to determine whether function words need to be inserted or deleted as compared to the SL. In addition, in Phase 2 cases of lexical ambiguity are resolved by selecting one lemma from each set of possible translations. That way, the best combination of lemmas is found for a given context. Finally, a token generator transforms TL lemmas into tokens.

A major objective of the PRESEMT project is to develop an MT system that can be easily extended to new language pairs. To this end the PRESEMT project uses readily available linguistic resources as far as possible and avoids the costly development of specialised linguistic resources and tools. Such tools include statistical taggers and chunkers that provide shallow linguistic structures.

3. Corpus comparison

As argued in Kilgarriff (2001), so long as we lack a systematic account of how one corpus relates to another, both corpus linguistics and corpus-based computational linguistics fall short of scientific standards. While that was as true when that work was done, in the 1990s, as it is now, it was perhaps forgivable then, since there were few corpora available so, in practice, scientists found themselves obliged to use whatever corpus (of the right language and, to some approximation, the right text type) was available. Now we can build corpora to order, automatically, from the web, so the question "how does this corpus relates to others I might use (of the same language) becomes critical. In PRESEMT we are following three strategies for addressing this question: Quantitative comparison, qualitative comparison, and evaluation (which we shall be reporting on later).

3.1 Qualitative comparison

Given two corpora, it has long been acknowledged that one way to get a sense of the differences between them is to look at the keywords of each *vs.* the other (see e.g. Hofland and Johanssen 1982). There has been debate on what statistics are most suitable for identifying keywords, and in Kilgarriff (2009) we make the case for:

- Normalising the frequency of each word in each corpus to a per-million figure
- Adding a parameter *k* to all normalised frequencies
- For each word, finding the ratio between the adjusted normalised frequencies in the two corpora.

The words with the highest ratio are then the keywords of corpus 1 *vs.* corpus 2, and those with the lowest are the keywords of corpus 2 *vs.* corpus 1. There are two advantages to adding *k* before taking the ratio: firstly, it allows us to take a ratio even when a word is absent in one of the corpora; and secondly, it allows us to vary *k* according to the focus of our research. A low value of *k* will tend to give lexical keywords, a higher value give more higher-frequency keywords, usually including grammatical words.

Then we can compare two corpora qualitatively by looking at the keywords of each *vs.* the other. It is usually possible to make some general statements about how the text type of each corpus differs from the text type of the other, by looking at the two lists of 100, or 200, keywords.

3.2 Quantitative comparison

Kilgarriff (2001) shows that a corpus distance measure based on frequency differences of the 500 commonest corpora work well to distinguish more, and less, similar text types. Within PRESEMT we have implemented a version of the 2001 measure within the Sketch Engine http://www.sketchengine.co.uk) so making it possible for researchers to classify which, of a set of three or more corpora for a language, are more similar and which are less so. Whereas the earlier work used a measure based on the chi-square statistic, we now use a variant of the same measure we use for keywords (with *k*=100, and taking the ratio by always dividing the higher number by the lower). We found this variant to be as precise as the one reported on before, and it is convenient to use a method consistent with keyword lists. The display we get for five well-known corpora of English is shown in Table 1.

	BASE	BAWE	BNC	Brown	BrownF
BASE		3.28	2.77	3.11	2.82
BAWE			2.15	2.21	2.09
BNC				1.59	1.32
Brown					1.47
BrownF					

Table 1: Distances between five well-known corpora of English: British Academic Spoken English (BASE), British Academic Written English (BAWE), the British National Corpus (BNC), the Brown corpus, and six 'Brown Family' corpora: Brown, LOB, FROWN, FLOB, BLOB, BE06.

The scores are 'average ratios', always guaranteed to be one (representing identical text types) or more. We can immediately see a cluster of the three corpora aiming at representativeness (BNC, Brown, Brown-Family), with the BASE, comprising spoken material, being the further-out outlier, and BAWE still an outlier but less different. We also note that Brown-family is slightly more similar to the BNC than it is to Brown, even though Brown is one of its component parts. This is perhaps because two thirds of Brown-family is British English, like the BNC, whereas Brown is entirely American.

Any user of the Sketch Engine can use the interface to find how a corpus of their own is situated in relation to other corpora. The interface does not as time of writing give a heterogeneity score for each corpus (which is needed, in order to interpret distance scores correctly) but will shortly be upgraded to provide this information.

References

Dologlou, Y., Markantonatou, S., Tambouratzis, G., Yannoutsou, O., Fourla, S., Ioannou, N. (2003) Using Monolingual Corpora for Statistical Machine Translation: The METIS System. EAMT-CLAW'03 Workshop Proceedings, Dublin, 15-17 May, pp. 61-68.

Hofland, K. & S. Johansson. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Kilgarriff, A. 2001. Comparing Corpora. *Int Jnl Corpus Linguistics,* 6 (1), pp 97-133

Kilgarriff, A. 2009. Simple Maths for Keywords. *Proc. Int Conf Corpus Linguistics*, Liverpool, UK.

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

Markantonatou S., S. Sofianopoulos, O. Giannoutsou & M. Vassiliou 2009: Hybrid Machine Translation for Low- and Middle- Density Languages. Language Engineering for Lesser-Studied Languages,, pp. 243-274. IOS Press.