Abstract. This paper describes different pivot approaches to built SMT systems for language pairs with scarce parallel resources. The strategy is particularly interesting for Spain, a country with three official languages (Catalan, Basque, and Galician) besides Spanish, where it is difficult to find parallel corpora between two of the first three mentioned languages but it is relatively easy to collect it between Spanish and any of them. This characteristic, however, allow us to develop machine translation systems from major languages like English, to Catalan for instance, using Spanish as pivot. Such systems help these minority languages giving them global presence and promoting their use in content collaboration. We describe a English-Catalan baseline system built following the synthetic approach, we compare it with the transfer approach and comment about future enhancement that could be implemented for this language pair.

1 Motivation

Spain is a multilingual country with four official languages: Catalan, Euskera, Galician and Spanish. Catalan is spoken by 11.5 million people, Euskera by 1.2 million people, Galician by 3.2 million people and Spanish by 400 million people. Given the high number of Spanish speakers compared to the other languages, Spanish has much more linguistic and data resources.

The quantity of resources is relevant in statistical machine translation. The more parallel text we have, the better the translation quality. In order to face the lack of resources in translation, there are many research works on pivot approaches which consist on using a pivot language to perform a source to target translation [1] [2]. For example, in order to translate from Galician to Catalan, we could use Spanish as pivot language. There are much more resources in Galician-Spanish and Spanish-Catalan than between Galician and Catalan directly. The same could happen when interested in translating
Catalan, Euskera or Galician into English. In this work, we introduce a state-of-the-art
English-Catalan translation system recently built for the free web translator N-II\(^1\).

The main differences with the Catalan-English SMT system presented in [3] are
that in this paper we use an extended corpus and we propose to build a hybrid system
which uses an Ngram-based system for Catalan-Spanish and a phrase-based system
for Spanish-English. The Ngram-based system outperforms the phrase-based system
in Catalan-Spanish [4] while the opposite occurs for the case of Spanish-English [5].
Additionally, for the Catalan-Spanish system we are using a further competitive system
using rules and statistical features [6].

The remainder of this paper is organized as follows. Section 2 reports a brief de-
scription of the phrase-based and Ngram-based translation approaches. Section 3 presents
the pivot approaches used in this paper. Section 4 describes the English-Catalan SMT
system. Section 5 compares the pivot strategies in terms of translation quality and Sec-
tion 6 presents the most relevant conclusions.

2 Statistical Machine Translation approaches

As mentioned in the previous section, we are working with two SMT systems: the
phrase-based [7] and Ngram-based systems [8, 9], which are briefly described as fol-
lows.

2.1 Phrase-based

This approach to SMT performs the translation splitting the source sentence in seg-
mants and assigning to each segment a bilingual phrase from a phrase-table. Bilin-
gual phrases are translation units that contain source words and target words, e.g. \(<
unidad de traducción | translation unit >\), and have different scores associated to
them. These bilingual phrases are then selected to maximize a linear combination of
feature functions. Such strategy is known as the log-linear model [10] and it is formally
defined as:

\[
\hat{e} = \arg \max_e \left[ \sum_{m=1}^{M} \lambda_m h_m (e, f) \right]
\]

where \(h_m\) are different feature functions with weights \(\lambda_m\). The two main feature func-
tions are the translation model (TM) and the target language model (LM). Additional
models include POS target language models, lexical weights, word penalty and reorder-
ing models among others.

Moses [11] was used to build the phrase-based system.

2.2 Ngram-based

The base of the Ngram approach is the concept of tuple. Tuples are bilingual units with
consecutive words both on the source and target side that are consistent with the word

\(^1\) available at http://www.n-ii.org
alignment. They must provide a unique monotonic segmentation of the sentence pair and they cannot be inside another tuple in the same sentence. This unique segmentation allows us to see the translation model as a language model, where the language is composed of tuples instead of words. That way, the context used in the translation model is bilingual and implicitly works as a language model with bilingual context as well. In fact, while a language model is required in phrase-based and hierarchical phrase-based systems, in Ngram-based systems it is considered just an additional feature.

This alternative approach to a translation model defines the probability as:

$$P(f, e) = \prod_{n=1}^{N} P(\{(f, e)_n \mid (f, e)_{n-1}, \ldots, (f, e)_1\})$$

(2)

where \((f, e)_n\) is the n-th tuple of hypothesis \(e\) for the source sentence \(f\).

As additional features, we used a Part-Of-Speech (POS) language model for the target side and a target word bonus model.

We used the open source decoder MARIE [12] to build the Ngram-based system.

3 Pivot Approaches

The best approaches to build a SMT system through a pivot language are: the cascade system, also known as the transfer approach and the pseudo-corpus or synthetic approach. Other pivot approaches do not outperform these two [13] [14]. The cascade and the pseudo-corpus approaches have been evaluated and compared in works such as [3, 1, 15]. Consistently, both works have shown that the pseudo-corpus approach is the best performing strategy.

3.1 Cascade or transfer method

This approach considers the language pairs source-pivot and pivot-target independently. It consists in training and tuning two different SMT systems and combine them in a two-step process: first, we translate a source sentence using the source-pivot system; then, we use the resulting sentence as input for the pivot-target translation. A common variation for this strategy presented in [16] considers a \(n\)-best output instead of the single-best during the first translation and then produce a \(m\)-best translation in the last step. At the end, \(mn\)-best hypotheses are produced, which are reranked by using Minimum Bayes Risk (MBR) [17], allowing the introduction of additional features such as new language models.

3.2 Pseudo-corpus or synthetic approach

Instead of considering the two language pairs independently, this approach produces a single source-target SMT system. Assuming we have a source-pivot and a pivot-target parallel corpus, we build and tuned a pivot-target SMT system and we use it to translate the pivot part from the source-pivot corpus. This results in a source-target synthetic corpus (hence the name) which is finally used to build the source-target SMT system.
For the tuning process, we could also use a synthetic development corpus but an actual source-target corpus is preferred, if possible. A simple variation for this approach is to build a pivot-source SMT system in order to translate the pivot part of the pivot-target corpus, and use the resulting source-target synthetic corpus to build the final system.

4 Building an English-Catalan SMT using Spanish as pivot

We present an English-Catalan SMT baseline system, using Spanish as the pivot language. In this case, the parallel corpus available for the Catalan-Spanish language pair was provided by the bilingual newspaper “El Periódico”\(^2\) and the English-Spanish corresponds to the train corpora provided during the 2010 WMT’s translation task\(^3\), i.e. Europarl and News Commentary. We followed the synthetic approach described before to build the final system. Therefore, the Spanish part from the WMT Corpus was translated into Catalan and an English-Catalan phrase-based SMT system was built using the resulting synthetic corpus. Table 1 shows a summary of the statistics of both corpora. We also used the Catalan-Spanish baseline together with the Spanish-English baseline system presented in the 2010’s WMT [18] to build the other direction and compare the different approaches in it.

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Catalan</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training sents.</td>
<td>4.6M</td>
<td>4.6M</td>
</tr>
<tr>
<td>Running words</td>
<td>96.94M</td>
<td>96.86M</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1.28M</td>
<td>1.23M</td>
</tr>
<tr>
<td>Development sents.</td>
<td>1966</td>
<td>1966</td>
</tr>
<tr>
<td>Running words</td>
<td>46765</td>
<td>44667</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>9132</td>
<td>9426</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Spanish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training sents.</td>
<td>1.18M</td>
<td>1.18M</td>
</tr>
<tr>
<td>Running words</td>
<td>26.45M</td>
<td>25.29M</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>118073</td>
<td>89248</td>
</tr>
<tr>
<td>Development sents.</td>
<td>1729</td>
<td>1729</td>
</tr>
<tr>
<td>Running words</td>
<td>37092</td>
<td>34774</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>7025</td>
<td>6199</td>
</tr>
<tr>
<td>Test sents.</td>
<td>2525</td>
<td>2525</td>
</tr>
<tr>
<td>Running words</td>
<td>69565</td>
<td>65595</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>10539</td>
<td>8907</td>
</tr>
</tbody>
</table>

Table 1. Catalan-Spanish and Spanish-English corpora (\(M\) stands for Millions)

\(^2\) [http://www.elperiodico.es](http://www.elperiodico.es)

\(^3\) [http://www.statmt.org/wmt10/translation-task.html](http://www.statmt.org/wmt10/translation-task.html)
4.1 Spanish-Catalan baseline system

As mentioned before, the Spanish-Catalan SMT system (named N-II) is based on the corpus provided by the bilingual newspaper “El Periódico”. It is a Ngram-based SMT system that includes several improvements specific to the language pair: a homonym disambiguation for the Catalan verb ‘soler’ and Catalan possessives, special consideration for pronominal clitics, upper-case words and the Catalan apostrophe, gender concordance, numbers and time categorization and text processing for common mistakes found when writing in Catalan. The full description can be found in [6].

4.2 English-Catalan system description

Once obtained the Catalan translation from the Spanish section of the WMT corpus, a phrase-based SMT system was built using Moses as the decoder. Apart from the baseline pipeline, the system also includes a POS target language model computed with TnT [19], numbers and time categorization similar to N-II and the parallel corpus was aligned considering the Catalan lemmas computed with Freeling [20] and the English stems of words obtained with Snowball.

5 Results

Table 2 shows the BLEU score of the cascade and pseudo-corpus approaches in both directions. The test set was the one provided as internal test set during the WMT translation task. It is also important to mention that the score was computed using one reference.

<table>
<thead>
<tr>
<th>Pivot approach</th>
<th>Direction</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cascade</td>
<td>cat-eng</td>
<td>21.63</td>
</tr>
<tr>
<td>Cascade</td>
<td>eng-cat</td>
<td>24.29</td>
</tr>
<tr>
<td>Pseudo-corpus</td>
<td>cat-eng</td>
<td>23.19</td>
</tr>
<tr>
<td>Pseudo-corpus</td>
<td>eng-cat</td>
<td>26.97</td>
</tr>
</tbody>
</table>

Table 2. English-Catalan results

The final quality of the Catalan-English system is determined by the quality of the Spanish-English corpus, whose baseline has a BLEU around 24 [18]. The Catalan-Spanish baseline has a BLEU around 80 [4]. Also there is a negative effect given the difference in domain between the Catalan-Spanish corpus (a regional newspaper) and Spanish-English corpus (Europarl).

Using paired bootstrap resampling [21], we can see that for these systems, the Pseudo-corpus approach is better than Cascade with 95% statistical significance.

4 http://snowball.tartarus.org
6 Conclusions and further work

We have presented an English-Catalan SMT system built using Spanish as pivot language, given the scarce resources for English-Catalan.

Similarly to previous research work, we have seen here that, in the particular translation task under consideration, the pseudo-corpus approach constitutes the best strategy for pivot translation. Although the cascade approach clearly performs worse than the pseudo-corpus approach, it could be also beneficial to consider a system combination between these two strategies to further boost the quality of the translations.

Further work should focus on building Spanish-pivot systems between all the official languages and English, as well as among them. The similarities between the languages (except Basque) and the availability of parallel corpora between Spanish and the others encourage the approach.

Acknowledgment

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement 247762 (FAUST) and from the Spanish Ministry of Science and Innovation through the Juan de la Cierva research program and the Buceador project (TEC2009-14094-C04-01).

References