In this paper we present the construction process of a web corpus of Catalan built from the content of the .cat top-level domain. For collecting and processing data we use the Brno pipeline with the spiderling crawler and its accompanying tools. To the best of our knowledge the corpus represents the largest existing corpus of Catalan containing 687 million words, which is a significant increase given that until now the biggest corpus of Catalan, CuCWeb, counts 166 million words. We evaluate the resulting resource on the tasks of language modeling and statistical machine translation (SMT) by calculating LM perplexity and incorporating the LM in the SMT pipeline. We compare language models trained on different subsets of the resource with those trained on the Catalan Wikipedia and the target side of the parallel data used to train the SMT system.

Keywords: web corpus, Catalan, language modeling, perplexity, machine translation

1. Introduction

The approach of building large corpora from the web has become the mainstream approach in the last decade. It has gained momentum with the WackY initiative inside which web corpora of English, German and Italian were constructed (Baroni et al., 2009). Other initiatives followed, such as the CoW collection of web corpora (Schäfer and Bildhauer, 2012). Recently a full pipeline for building top-level domain (TLD) web corpora has emerged – the Brno pipeline (Suchomel and Pomikálek, 2012) as part of which multiple tools for crawling, encoding detection, content extraction and near-deduplication were published.

In this paper we present the procedure of building a web corpus of Catalan from documents published on the .cat top-level domain. Until now the largest Catalan corpus was the CuCWeb corpus (Boleda et al., 2006) which is built from the web and contains 166 million words. It was built as a by-product of constructing a Spanish web corpus by crawling data that is hosted in Spain regardless of the TLD. Each document in the corpus was classified by language and the subset of documents identified as Catalan was separately published as CuCWeb. The data for the corpus was collected in September and October 2004. Since then the situation regarding the presence of Catalan on the web has significantly changed with the emergence of the sponsored TLD .cat being approved in September 2005. Our approach to building the web corpus of Catalan exploits exactly that fact by crawling that TLD.

We construct the corpus with the aforementioned Brno pipeline, mostly with default settings. We do focus on the various levels of duplicate removal – physical, near-document and near-paragraph – and inspect the impact of each level on the number of unique sentences used for language modeling. We inspect the language models on two levels – by checking their perplexity and by using them in a statistical machine translation (SMT) task to translate from Spanish to Catalan. We compare the data collection crawled with a small in-domain corpus from the SMT task and with a corpus constructed from the Catalan Wikipedia dump.

The remainder of this paper is structured as follows: in Section 2 we describe the process of constructing various versions of the corpora and give a comparison between them while in Section 3 we apply the constructed corpora to the language modeling task for SMT. In Section 4 we provide a conclusion and describe our further steps.

2. Corpus construction

This section details the procedure of building caWaC, a web corpus of Catalan from the .cat top-level domain. We have built the caWaC corpus with the Brno pipeline (Suchomel and Pomikálek, 2012) which consists of the following tools:

- the spiderling crawler used for collecting HTML documents from the web,
- the chared encoding detector that guesses the correct encoding of each crawled document,
- the justext content extractor which returns the linguistically relevant text from the document,
The trigram language identification algorithm and
the onion near-duplicate removal tool[1] which removes
documents or paragraphs that have a defined n-gram overlap with already seen data.

The spiderling crawler requires as input a list of seed URLs, the number of threads to use for document processing, the maximum duration of the crawl and the size ratio threshold that is used to stop the crawling of a low-yield-rate domain. As seed URLs we used the list of .cat domains which are in the top 1 million sites globally by traffic; there were 252 such domains[2] at our time of retrieval. We used 16 threads for document processing, crawled for 21 days and used the lower predefined size ratio threshold recommended for smaller languages. The size ratio threshold is used to stop crawling domains with low yield rate regarding the final content. The size ratio is calculated as the ratio between the amount of final data after post-processing (except near-duplicate removal) and the amount of downloaded data. A domain is dropped from further crawling once this ratio falls below the predefined threshold.

For encoding detection with chared we used the predefined list of function words for Catalan distributed with the tool. Similarly, for content extraction, we used the predefined list of function words for Catalan distributed with justext.

Language identification was performed with the trigram method, which calculates a character trigram vector for a controlled sample of the language sought and a trigram vector for each document after content extraction. It calculates the cosine similarity, transformed to a distance measure, between those two vectors and discards the document if the distance is above a specific threshold. We defined that threshold to be 0.8.

The language sample used to perform language identification was built with a short initial crawl for which language identification was performed with a clean sample of 20,000 words from the Catalan Wikipedia dump. For building the final language sample we selected documents from that initial crawl with the character trigram distance below 0.2 and kept paragraphs longer than 50 characters. With this procedure we obtained a clean sample of 3.6 million words of Catalan.

The crawl was run from 2013-08-30 to 2013-09-13. After finishing this two-week crawl and performing exact document-level deduplication of the corpus, we tokenised and sentence split the corpus with the Freeling toolkit [Padró and Stanilovsky, 2012]. The tokenized corpus was near-duplicated with onion with default settings, both on paragraph and on document level.

Basic statistics about the corpora obtained with different levels of duplicate removal are given in Table 1. We can observe that near-duplicate removal on the document-level leaves 57% of the original tokens, while after near-duplicate removal on the paragraph level only 28% of the original tokens remain. Namely, we remove additional 50% of the tokens that survived the first stage of near-duplicate removal. Our plan for the next section is to inspect how each of these deduplication approaches influences the performance of the resulting language model in terms of perplexity and when used in a translation task.

<table>
<thead>
<tr>
<th>Deduplication</th>
<th>Level</th>
<th>Type</th>
<th># of sent</th>
<th># of tok</th>
</tr>
</thead>
<tbody>
<tr>
<td>document</td>
<td>exact</td>
<td></td>
<td>53,149,225</td>
<td>1,378,350,380</td>
</tr>
<tr>
<td>document</td>
<td>near</td>
<td></td>
<td>31,122,527</td>
<td>779,086,559</td>
</tr>
<tr>
<td>paragraph</td>
<td>near</td>
<td></td>
<td>15,068,847</td>
<td>384,548,042</td>
</tr>
</tbody>
</table>

Table 1: Number of sentences and tokens on various levels of duplicate removal

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of unique sents</th>
<th># of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>6,891,483</td>
<td>136,887,790</td>
</tr>
<tr>
<td>caWaC.ed</td>
<td>4,627,411</td>
<td>137,393,049</td>
</tr>
<tr>
<td>caWaC.dnd</td>
<td>4,696,776</td>
<td>137,622,883</td>
</tr>
<tr>
<td>caWaC.pnd</td>
<td>4,818,837</td>
<td>137,712,738</td>
</tr>
<tr>
<td>caWaC.ed</td>
<td>24,745,986</td>
<td>733,974,675</td>
</tr>
<tr>
<td>caWaC.dnd</td>
<td>21,486,116</td>
<td>625,569,184</td>
</tr>
<tr>
<td>caWaC.pnd</td>
<td>13,403,229</td>
<td>382,984,233</td>
</tr>
</tbody>
</table>

Table 2: Number of sentences and tokens in the datasets used for language modeling

3. Evaluation

In order to evaluate caWaC, we perform two types of evaluation, in the tasks of language modeling (LM) and statistical machine translation (SMT).

We perform all language modeling on datasets that contain unique sentences only, as this is the standard approach in SMT.

We build language models on the data from different versions of caWaC regarding the level of deduplication. We differentiate the following three versions:

1. exact duplicate removal on the document level (caWaC.ed),
2. near-duplicate removal on the document level (caWaC.dnd) and
3. near-duplicate removal on paragraph level (caWaC.pnd).

In addition, we want to compare caWaC to other possible sources for LM data, in our case Wikipedia. The Wikipedia corpus was prepared with a standard clean-up script from a dump of the Catalan Wikipedia retrieved on 2013-09-14. Finally, for a fair comparison with Wikipedia we not only consider the full sets from caWaC but also subsets (ws) whose size is comparable, i.e. they contain the same number of bytes as that of the Wikipedia dataset. It is important to stress that the caWaC subsets of the same size as Wikipedia were built from datasets containing unique sentences only to eliminate the possible impact of unique sentence density in various datasets.

The sizes of the LM datasets (number of unique sentences and number of running tokens) are given in Table 2. At this point we can inspect the impact of different levels of duplicate removal on the percentage of unique sentences
Table 3: Statistics of the development and test sets

<table>
<thead>
<tr>
<th>Set</th>
<th>Type</th>
<th># of tokens</th>
<th>Voc size</th>
<th>OOVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>dev</td>
<td>21,357</td>
<td>6,172</td>
<td>0.7%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>21,590</td>
<td>6,316</td>
<td>0.8%</td>
</tr>
<tr>
<td>Tatoeba</td>
<td>dev</td>
<td>8,252</td>
<td>2,200</td>
<td>0.9%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>8,962</td>
<td>2,210</td>
<td>1.1%</td>
</tr>
<tr>
<td>EUBookshop</td>
<td>dev</td>
<td>26,001</td>
<td>4,764</td>
<td>4.2%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>24,874</td>
<td>4,471</td>
<td>4.3%</td>
</tr>
<tr>
<td>Novel</td>
<td>dev</td>
<td>22,412</td>
<td>4,808</td>
<td>5.7%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>17,335</td>
<td>3,536</td>
<td>6.4%</td>
</tr>
</tbody>
</table>

We compare the perplexities obtained by LMs built on different versions of caWaC, an LM built on the data from a Wikipedia dump and an LM built on the target side of the parallel corpus used in SMT (news).

Perplexities are computed for the four test sets used in the SMT task, thus we assess the performance in terms of perplexity of caWaC-based LMs for data of the domain that is covered in the baseline (news) and for other domains (Tatoeba, EUBookshop and novel).

We consider the case where LMs are built and tested on news data as in-domain while all other cases belong to the out-of-domain scenario.

For the baseline LM we take the Wikipedia dataset as we consider it to be, at least in the out-of-domain scenario, a worthier opponent to the caWaC datasets than the small news dataset. The results on 8 different datasets and 4 different test sets are given in Table 4.

The in-domain scenario (system News, test News) expectantly outperforms the comparable out-of-domain scenarios (other systems, test News). From the results in the out-of-domain scenario we can draw the following conclusions:

- caWaC datasets of comparable size to the Wiki dataset outperform the Wiki dataset significantly on all test sets proving web corpora are more diverse, these results are in accordance to our previous comparison of that resource types in the task of identifying false friends via distributional methods (Ljubešić and Fiser, 2013)
- full-blown caWaC datasets perform better than the wiki-size caWaC datasets showing the usefulness of the large amount of information we collected.
- the biggest caWaC dataset, caWaC.ed performs consistently slightly worse than its document-level and paragraph-level near-deduplicated counterparts, the positive impact of near-duplicate removal could be explained through elimination of remains of boilerplate that distort the LM statistics, and
- there is no decisive winner when comparing document-level and paragraph-level near-deduplicated datasets, but, given the smaller size of paragraph-level near-deduplicated datasets, they should be given advantage over datasets where near-duplicate removal was performed on the document-level.

3.2. Machine Translation

We experiment with an SMT system trained on the news domain for the Spanish to Catalan language direction. Again, we want to inspect the contribution of caWaC as a LM to the performance of the SMT system in our two scenarios: in-domain where the parallel and test data are both from the news domain, and out-of-domain where the test data are from one of the three remaining domains.

The baseline MT system is trained on a parallel corpus from the news domain, made up of 10 years of bilingual articles.
The corpus contains 633,257 unique sentence pairs, and the target side is used to build the LM. For training we use the subset of sentences with 1 to 80 tokens: 629,375.

The Moses toolkit (Koehn et al., 2007) is used to train a phrase based SMT system with the parallel data previously introduced. Tuning is carried out with MERT (Och, 2003).

We computed also TER (Snover et al., 2006) and METEOR (Lavie and Denkowski, 2009) scores, due to the similarity of the trends for those metrics they are omitted.

Finally, we would like to mention that the caWaC corpus in language modeling and SMT. In both cases the addition of data crawled from Wikipedia.

It is worth to note that although the paper has dealt with a specific study for Catalan, our approach to acquire monolingual data from the web and the application of this data to improve the performance of tasks such as language modeling and SMT should be applicable to any other languages, specially those for which no large monolingual corpora are publicly available yet.

As future work, we envisage two lines of research. On the one hand, we would like to experiment with unsupervised methods for corpus filtering based on identifying outliers regarding character n-gram probabilities as described in (Ljubešić and Klubička, 2014). On the other hand, in this work we applied LMs built on crawled data in SMT for a language pair of closely-related languages. We envisage that the addition of such LMs could lead to bigger improvements when applied to translation between more distant languages, e.g. English to Catalan.

4. Conclusions

This paper has presented caWaC, a corpus for Catalan crawled from the web. To the best of our knowledge this is the largest corpus of Catalan available and the first one to be freely available.

We have detailed the acquisition procedure, using the Brno crawling pipeline. We have then applied the acquired corpus in language modeling and SMT. In both cases caWaC results in better performance when compared to data crawled from Wikipedia.

Surprisingly, given our previous results with language modeling (cf. Section 5.1), using the whole caWaC datasets does not result in an improvement over the caWaC subsets of Wikipedia size, even if the former are up to 5 times bigger (see Table 2). The different values of deduplication strategies (ed, dnd and pnd) do not seem to have much of an impact on this close language pair, as the differences between caWaC systems are very small.

4. Conclusions

This paper has presented caWaC, a corpus for Catalan crawled from the web. To the best of our knowledge this is the largest corpus of Catalan available and the first one to be freely available.

We have detailed the acquisition procedure, using the Brno crawling pipeline. We have then applied the acquired corpus in language modeling and SMT. In both cases caWaC results in better performance when compared to data crawled from Wikipedia.

It is worth to note that although the paper has dealt with a specific study for Catalan, our approach to acquire monolingual data from the web and the application of this data to improve the performance of tasks such as language modeling and SMT should be applicable to any other languages, specially those for which no large monolingual corpora are publicly available yet.

As future work, we envisage two lines of research. On the one hand, we would like to experiment with unsupervised methods for corpus filtering based on identifying outliers regarding character n-gram probabilities as described in (Ljubešić and Klubička, 2014). On the other hand, in this work we applied LMs built on crawled data in SMT for a language pair of closely-related languages. We envisage that the addition of such LMs could lead to bigger improvements when applied to translation between more distant languages, e.g. English to Catalan.

Finally, we would like to mention that the caWaC corpus is freely available for download[11] for IPR reasons in sentence-scrambled format, under the CC-BY-SA license[12].

[13] http://creativecommons.org/licenses/by-sa/3.0/
5. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

6. References


