caWaC – A web corpus of Catalan and its application to language modeling and machine translation

Nikola Ljubešić,* Antonio Toral†

* Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
nikola.ljubesic@ffzg.hr

† School of Computing
Dublin City University
Dublin 9, Ireland
atoral@computing.dcu.ie

Abstract

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[1] used for collecting HTML documents from the web,
- the *chared* encoding detector[2] that guesses the correct encoding of each crawled document,
- the *justext* content extractor[3] which returns the linguistically relevant text from the document,

[1] `http://nlp.fi.muni.cz/trac/spiderling`
[2] `https://code.google.com/p/chared/`
[3] `http://code.google.com/p/justext/`

- the *trigram* language identification algorithm and
- the *onion* near-duplicate removal tool [4] 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[5] 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 prebuilt model 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-deduplicated 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.

[4] http://code.google.com/p/onion/

[5] https://domaintyper.com/top-websites/most-popular-websites-with-cat-domain

Deduplication			
Level	Type	# of sent	# of tok
document	exact	53,149,225	1,378,350,380
document	near	31,122,527	779,086,559
paragraph	near	15,068,847	384,548,042

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

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

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

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

Table 3: Statistics of the development and test sets

in each dataset by calculating the ratio between the number of sentences after and before sentence deduplication from Tables 1 and 2. This ratio is 0.466 for the corpus after exact duplicate removal on the document level (meaning that 46.6% of sentences are unique in this version of the corpus), 0.69 after near-duplicate removal on the document level and 0.889 after near-duplicate removal on the paragraph level. This shows, as expected, that harsher duplicate removal does increase the percentage of unique sentences. On the other hand, after the two levels of near-duplicate removal, only 54% of unique sentences from the initial corpus are still present. This is quite a drastic loss of unique sentences and we are thus interested in observing what the impact will be on our chosen tasks.

Each of these LMs is tested by (i) calculating the perplexity given a test set and (ii) applying it to the task of SMT, by incorporating it to the SMT pipeline as the LM of the translation system. On both tasks we would like to assess the contribution of a specific LM in two scenarios:

- In-domain. The LM data (for the perplexity experiment) or the parallel training data (for the SMT experiment) belong to the same domain as the test data,
- Out-of-domain. Test data are from a different domain than the LM data or the parallel training data.

All the LMs are built with the IRSTLM toolkit (Federico et al., 2008), they consider n-grams up to order 5 and they are smoothed using a simplified version of the modified Kneser-Ney method (Chen and Goodman, 1996).

The LMs are evaluated for both tasks on four test sets. One of them is in-domain (news), while the other three are out-of-domain and belong to different genres: official publications from the European Union (EUbookshop[6]), sentences for language learners (Tatoeba[7]) and literature (a Spanish contemporary best seller novel). The SMT systems are tuned on development data extracted from the same dataset as the test data. Each development and test set are made of 1,000 sentences. Perplexity is calculated on the test sets only. Statistics of the development and test sets (number of tokens and vocabulary size of the source side and percentage of out-of-vocabulary words (OOVs) with respect to the parallel corpus used for training the SMT system) can be found in Table 3.

[6] http://opus.lingfil.uu.se/EUbookshop.php
[7] http://opus.lingfil.uu.se/Tatoeba.php

3.1. Perplexity

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) expectedly 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 Fišer, 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

System	News		EUbookshop		Tatoeba		Novel	
	PPL	Δ%	PPL	Δ%	PPL	Δ%	PPL	Δ%
Wiki	386.86		274.26		327.11		458.84	
News	**134.19**	**-65.31%**	386.76	41.02%	285.28	-12.79%	850.97	85.46%
caWaC.ed.ws	224.32	-42.02%	206.1	-24.85%	135.02	-58.72%	267.55	-41.69%
caWaC.dnd.ws	221.26	-42.81%	201.53	-26.52%	128.31	-60.77%	253.14	-44.83%
caWaC.pnd.ws	212.78	-45.00%	199.34	-27.32%	116.04	-64.53%	227.79	-50.36%
caWaC.ed	187.69	-51.48%	177.81	-35.17%	110.69	-66.16%	227.89	-50.33%
caWaC.dnd	**185.01**	**-52.18%**	**173.35**	**-36.79%**	106.77	-67.36%	218.34	-52.41%
caWaC.pnd	187.05	-51.65%	177.57	-35.25%	**101.19**	**-69.07%**	**202.14**	**-55.95%**

Table 4: Perplexity results

System	News		EUbookshop		Tatoeba		Novel	
	BLEU	Δ%	BLEU	Δ%	BLEU	Δ%	BLEU	Δ%
baseline	0.8465		0.4111		0.6449		0.6347	
Wiki	0.8427	-0.45%	0.4150	0.95%	0.6518	1.07%	0.6389	0.66%
caWaC.ed.ws	0.8479	0.17%	0.4165	1.31%	0.6651	3.13%	0.6444	1.53%
caWaC.dnd.ws	0.8475	0.12%	0.4131	0.49%	0.6586	2.12%	0.6410	0.99%
caWaC.pnd.ws	0.8474	0.11%	**0.4169**	**1.41%**	**0.6731**	**4.37%**	**0.6454**	**1.69%**
caWaC.ed	0.8484	0.22%	0.4153	1.02%	0.6704	3.95%	0.6443	1.51%
caWaC.dnd	0.8482	0.20%	0.4154	1.05%	0.6714	4.11%	0.6435	1.39%
caWaC.pnd	**0.8486**	**0.25%**	0.4162	1.24%	0.6698	3.86%	0.6439	1.45%

Table 5: SMT results

(Spanish–Catalan) from "El Periódico de Catalunya".[8] 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). Table 5 shows the BLEU (Papineni et al., 2002) scores[9] on the four test sets for the baseline system and for the systems which add to the LM the datasets described in Table 2.

The addition of data acquired from the web results in better performance, the only exception being using Wikipedia's for the in-domain test set (-0.45%). Using data from caWaC for the LM allows the system to achieve higher scores than using Wikipedia's for all the four test sets. Our hypothesis for this is that the contents from the web corpus have more variety than Wikipedia's.

While caWaC improves the results on all the test sets, it provides a bigger improvement on out-of-domain datasets (ranging from 1.41% to 4.37%, relative) than on the in-domain test set (0.25%). We thus conclude that the addition of caWaC to the LM is especially useful for translating text in domains for which parallel training data is not available, and thus we deem it useful for domain adaptation of SMT (Bertoldi and Federico, 2009).

Surprisingly, given our previous results with language modeling (cf. Section 3.1.), 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[10], for IPR reasons in sentence-scrambled format, under the CC-BY-SA license[11].

[8] http://catalog.elra.info/product_info.php?products_id=1122

[9] 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.

[10] http://nlp.ffzg.hr/resources/corpora/cawac/

[11] 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

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, pages 209–226.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 182–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

G. Boleda, S. Bott, R. Meza, C. Castillo, T. Badia, V. Lopez, Grup De Lingüística Computacional, and Cátedra Telefónica De Producción Multimedia. 2006. Cucweb: a catalan corpus built from the web. In *Proceedings of the Second Workshop on the Web as a Corpus at EACL'06*.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621. ISCA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115, September.

Nikola Ljubešić and Filip Klubička. 2014. bs,hr,srwac – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of WAC-9 workshop*. Association for Computational Linguistics.

Nikola Ljubešić and Darja Fišer. 2013. Identifying false friends between closely related languages. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 69–77, Sofia, Bulgaria, August. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *In Proceedings of the Association for Machine Transaltion in the Americas (AMTA 2006*.

Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In Serge Sharoff Adam Kilgarriff, editor, *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43, Lyon.