Hybrid Selection of Language Model Training Data Using Linguistic Information and Perplexity

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

Abstract

We explore the selection of training data for language models using perplexity. We introduce three novel models that make use of linguistic information and evaluate them on three different corpora and two languages. In four out of the six scenarios a linguistically motivated method outperforms the purely statistical state-of-the-art approach. Finally, a method which combines surface forms and the linguistically motivated methods outperforms the baseline in all the scenarios, selecting data whose perplexity is between 3.49% and 8.17% (depending on the corpus and language) lower than that of the baseline.

1 Introduction

Language models (LMs) are a fundamental piece in statistical applications that produce natural language text, such as machine translation and speech recognition. In order to perform optimally, a LM should be trained on data from the same domain as the data that it will be applied to. This poses a problem, because in the majority of applications, the amount of domain-specific data is limited.

A popular strand of research in recent years to tackle this problem is that of training data selection. Given a limited domain-specific corpus and a larger non-domain-specific corpus, the task consists on finding suitable data for the specific domain in the non-domain-specific corpus. The underlying assumption is that a non-domain-specific corpus, if broad enough, contains sentences similar to a domain-specific corpus, which therefore, would be useful for training models for that domain.

This paper focuses on the approach that uses perplexity for the selection of training data. The first works in this regard (Gao et al., 2002; Lin et al., 1997) use the perplexity according to a domain-specific LM to rank the text segments (e.g. sentences) of non-domain-specific corpora. The text segments with perplexity less than a given threshold are selected.

A more recent method, which can be considered the state-of-the-art, is Moore-Lewis (Moore and Lewis, 2010). It considers not only the cross-entropy[1] according to the domain-specific LM but also the cross-entropy according to a LM built on a random subset (equal in size to the domain-specific corpus) of the non-domain-specific corpus. The additional use of a LM from the non-domain-specific corpus allows to select a subset of the non-domain-specific corpus which is better (the perplexity of a test set of the specific domain has lower perplexity on a LM trained on this subset) and smaller compared to the previous approaches. The experiment was carried out for English, using Europarl (Koehn, 2005) as the domain-specific corpus and LDC Gigaword[2] as the non-domain-specific one.

In this paper we study whether the use of two types of linguistic knowledge (lemmas and named entities) can contribute to obtain better results within the perplexity-based approach.

2 Methodology

We explore the use of linguistic information for the selection of data to train domain-specific LMs from non-domain-specific corpora. Our hypothesis is that ranking by perplexity on n-grams that represent linguistic patterns (rather than n-grams that represent surface forms) captures additional information, and thus may select valuable data that is not selected according solely to surface forms.

We use two types of linguistic information at

[1] note that using cross-entropy is equivalent to using perplexity since they are monotonically related.

[2] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2007T07

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 8–12,
Sofia, Bulgaria, August 8, 2013.

word level: lemmas and named entity categories. We experiment with the following models:

- Forms (hereafter f), uses surface forms. This model replicates the Moore-Lewis approach and is to be considered the baseline.
- Forms and named entities (hereafter fn), uses surface forms, with the exception of any word detected as a named entity, which is substituted by its type (e.g. person, organisation).
- Lemmas (hereafter l), uses lemmas.
- Lemmas and named entities (hereafter ln), uses lemmas, with the exception of any word detected as a named entity, which is substituted by its type.

A sample sentence, according to each of these models, follows:

```
f: I declare resumed the session of the
 European Parliament
fn: I declare resumed the session of the
 NP00O00
l: i declare resume the session of the
 european_parliament
ln: i declare resume the session of the
 NP00O00
```

Table 1 shows the number of n-grams on LMs built on the English side of News Commentary v8 (hereafter NC) for each of the models. Regarding 1-grams, compared to f, the substitution of named entities by their categories (fn) results in smaller vocabulary size (-24.79%). Similarly, the vocabulary is reduced for the models l (-8.39%) and ln (-44.18%). Although not a result in itself, this might be an indication that using linguistically motivated models could be useful to deal with data sparsity.

n	f	fn	l	ln
1	65076	48945	59619	36326
2	981077	847720	835825	702118
3	2624800	2382629	2447759	2212709
4	3633724	3412719	3523888	3325311
5	3929751	3780064	3856917	3749813

Table 1: Number of n-grams in LMs built using the different models

Our procedure follows that of the Moore-Lewis method. We build LMs for the domain-specific corpus and for a random subset of the non-domain-specific corpus of the same size (number of sentences) of the domain-specific corpus. Each sentence s in the non-domain-specific corpus is then scored according to equation 1 where $PP_I(s)$ is the perplexity of s according to the domain-specific LM and $PP_O(s)$ is the perplexity of s according to the non-domain-specific LM.

$$score(s) = PP_I(s) - PP_O(s) \quad (1)$$

We build LMs for the domain-specific and non-domain-specific corpora using the four models previously introduced. Then we rank the sentences of the non-domain-specific corpus for each of these models and keep the highest ranked sentences according to a threshold. Finally, we build a LM on the set of sentences selected[3] and compute the perplexity of the test set on this LM.

We also investigate the combination of the four models. The procedure is fairly straightforward: given the sentences selected by all the models for a given threshold, we iterate through these sentences following the ranking order and keeping all the distinct sentences selected until we obtain a set of sentences whose size is the one indicated by the threshold. I.e. we add to our distinct set of sentences first the top ranked sentence by each of the methods, then the sentence ranked second by each method, and so on.

3 Experiments

3.1 Setting

We use corpora from the translation task at WMT13.[4] Our domain-specific corpus is NC, and we carry out experiments with three non-domain-specific corpora: a subset of Common Crawl[5] (hereafter CC), Europarl version 7 (hereafter EU), and United Nations (Eisele and Chen, 2010) (hereafter UN). We use the test data from WMT12 (newstest2012) as our test set. We carry out experiments on two languages for which these corpora are available: English (referred to as "en" in tables) and Spanish ("es" in tables).

We test the methods on three very different non-domain-specific corpora, both in terms of the topics that they cover (text crawled from web in CC, parliamentary speeches in EU and official documents from United Nations in UN) and their size

[3] For the linguistic methods we replace the sentences selected (which contain lemmas and/or named entities) with the corresponding sentences in the original corpus (containing only word forms).

[4] http://www.statmt.org/wmt13/translation-task.html

[5] http://commoncrawl.org/

(around 2 million sentences both for CC and EU, and around 11 million for UN). This can be considered as a contribution of this paper since previous works such as Moore and Lewis (2010) and, more recently, Axelrod et al. (2011) test the Moore-Lewis method on only one non-domain-specific corpus: LDC Gigaword and an unpublished general-domain corpus, respectively.

All the LMs are built with IRSTLM 5.80.01 (Federico et al., 2008), use up to 5-grams and are smoothed using a simplified version of the improved Kneser-Ney method (Chen and Goodman, 1996). For lemmatisation and named entity recognition we use Freeling 3.0 (Padró and Stanilovsky, 2012). The corpora are tokenised and truecased using scripts from the Moses toolkit (Koehn et al., 2007).

3.2 Experiments with Different Models

Figures 1, 2 and 3 show the perplexities obtained by each method on different subsets selected from the English corpora CC, EU and UN, respectively. We obtain these subsets according to different thresholds, i.e. percentages of sentences selected from the non-domain-specific corpus. These are the first $\frac{1}{64}$ ranked sentences, $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ and 1.[6] Corresponding figures for Spanish are omitted due to the limited space available and also because the trends in those figures are very similar.

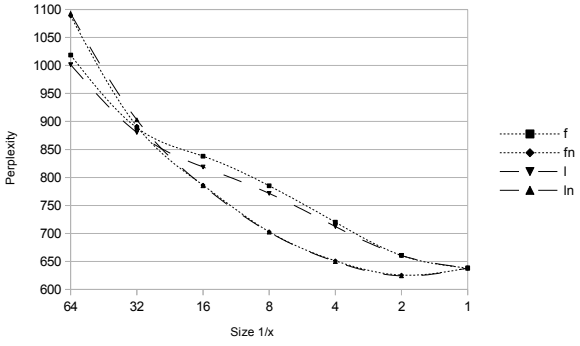

Figure 1: Results of the different methods on CC

In all the figures, the results are very similar regardless of the use of lemmas. The use of named entities, however, produces substantially different results. The models that do not use named entity categories obtain the best results for lower thresholds (up to $1/32$ for CC, and up to $1/16$ both for

[6] An additional threshold, $\frac{1}{128}$, is used for the United Nations corpus

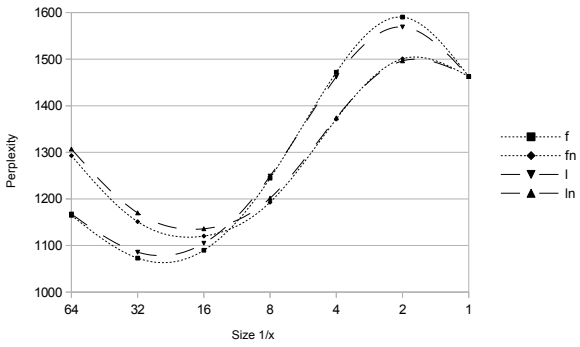

Figure 2: Results of the different methods on EU

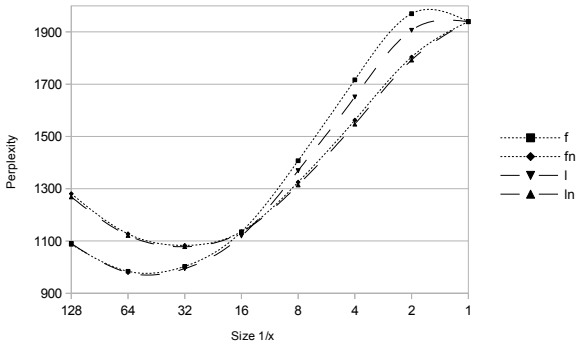

Figure 3: Results of the different methods on UN

EU and UN). If the best perplexity is obtained with a lower threshold than this (the case of EU, $1/32$, and UN, $1/64$), then methods that do not use named entities obtain the best result. However, if the optimal perplexity is obtained with a higher threshold (the case of CC, $1/2$), then using named entities yields the best result.

Table 2 presents the results for each model. For each scenario (corpus and language combination), we show the threshold for which the best result is obtained (column best). The perplexity obtained on data selected by each model is shown in the subsequent columns. For the linguistic methods, we also show the comparison of their performance to the baseline (as percentages, columns diff). The perplexity when using the full corpus is shown (column full) together with the comparison of this result to the best method (last column diff).

The results, as previously seen in Figures 1, 2 and 3, differ with respect to the corpus but follow similar trends across languages. For CC we obtain the best results using named entities. The model ln obtains the best result for English (5.54% lower

corpus	best	f	fn	diff	l	diff	ln	diff	full	diff
cc_en	1/2	660.77	625.62	-5.32	660.58	-0.03	**624.19**	-5.54	638.24	-2.20
eu_en	1/32	**1072.98**	1151.13	7.28	1085.66	1.18	1170.00	9.04	1462.61	-26.64
un_en	1/64	984.08	1127.55	14.58	**979.06**	-0.51	1121.45	13.96	1939.44	-49.52
cc_es	1/2	499.22	**480.17**	-3.82	498.93	-0.06	480.45	-3.76	481.96	-0.37
eu_es	1/16	**788.62**	813.32	3.13	801.50	1.63	825.13	4.63	960.06	-17.86
un_es	1/32	725.93	773.89	6.61	**723.37**	-0.35	771.25	6.24	1339.78	-46.01

Table 2: Results for the different models

perplexity than the baseline), while the model fn obtains the best result for Spanish (3.82%), although in both cases the difference between these two models is rather small.

For the other corpora, the best results are obtained without named entities. In the case of EU, the baseline obtains the best result, although the model l is not very far (1.18% higher perplexity for English and 1.63% for Spanish). This trend is reversed for UN, the model l obtaining the best scores but close to the baseline (-0.51%, -0.35%).

3.3 Experiments with the Combination of Models

Table 3 shows the perplexities obtained by the method that combines the four models (column comb) for the threshold that yielded the best result in each scenario (see Table 2), compares these results (column diff) to those obtained by the baseline (column f) and shows the percentage of sentences that this method inspected from the sentences selected by the individual methods (column perc).

corpus	f	comb	diff	perc
cc_en	660.77	**613.83**	-7.10	76.90
eu_en	1072.98	**1035.51**	-3.49	70.51
un_en	984.08	**908.47**	-7.68	74.58
cc_es	499.22	**478.87**	-4.08	74.61
eu_es	788.62	**748.22**	-5.12	68.05
un_es	725.93	**666.62**	-8.17	74.32

Table 3: Results of the combination method

The combination method outperforms the baseline and any of the individual linguistic models in all the scenarios. The perplexity obtained by combining the models is substantially lower than that obtained by the baseline (ranging from 3.49% to 8.17%). In all the scenarios, the combination method takes its sentences from roughly the top 70% sentences ranked by the individual methods.

4 Conclusions and Future Work

This paper has explored the use of linguistic information (lemmas and named entities) for the task of training data selection for LMs. We have introduced three linguistically motivated models, and compared them to the state-of-the-art method for perplexity-based data selection across three different corpora and two languages. In four out of these six scenarios a linguistically motivated method outperforms the state-of-the-art approach.

We have also presented a method which combines surface forms and the three linguistically motivated methods. This combination outperforms the baseline in all the scenarios, selecting data whose perplexity is between 3.49% and 8.17% (depending on the corpus and language) lower than that of the baseline.

Regarding future work, we have several plans. One interesting experiment would be to apply these models to a morphologically-rich language, to check if, as hypothesised, these models deal better with sparse data.

Another strand regards the application of these models to filter parallel corpora, e.g. following the extension of the Moore-Lewis method (Axelrod et al., 2011) or in combination with other methods which are deemed to be more suitable for parallel data, e.g. (Mansour et al., 2011).

We have used one type of linguistic information in each LM, but another possibility is to combine different pieces of linguistic information in a single LM, e.g. following a hybrid LM that uses words and tags, depending of the frequency of each type (Ruiz et al., 2012).

Given the fact that the best result is obtained with different models depending on the corpus, it would be worth to investigate whether given a new corpus, one could predict the best method to be applied and the threshold for which one could expect to obtain the minimum perplexity.

Acknowledgments

We would like to thank Raphaël Rubino for insightful conversations. The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreements PIAP-GA-2012-324414 and FP7-ICT-2011-296347.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

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.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for chinese. 1(1):3–33, March.

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.

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

Sung-Chien Lin, Chi-Lung Tsai, Lee-Feng Chien, Keh-Jiann Chen, and Lin-Shan Lee. 1997. Chinese language model adaptation based on document classification and multiple domain-specific language models. In George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas, editors, *EUROSPEECH*. ISCA.

Saab Mansour, Joern Wuebker, and Hermann Ney. 2011. Combining translation and language model scoring for domain-specific data filtering. In *International Workshop on Spoken Language Translation*, pages 222–229, San Francisco, California, USA, December.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, 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.

Nick Ruiz, Arianna Bisazza, Roldano Cattoni, and Marcello Federico. 2012. FBK's Machine Translation Systems for IWSLT 2012's TED Lectures. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*.