Language Dynamics and Capitalization using Maximum Entropy

Fernando Batista[a,b], **Nuno Mamede**[a,c] **and Isabel Trancoso**[a,c]
[a] L^2F – Spoken Language Systems Laboratory - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://www.l2f.inesc-id.pt/
[b] ISCTE – Instituto de Ciências do Trabalho e da Empresa, Portugal
[c] IST – Instituto Superior Técnico, Portugal.

Abstract

This paper studies the impact of written language variations and the way it affects the capitalization task over time. A discriminative approach, based on maximum entropy models, is proposed to perform capitalization, taking the language changes into consideration. The proposed method makes it possible to use large corpora for training. The evaluation is performed over newspaper corpora using different testing periods. The achieved results reveal a strong relation between the capitalization performance and the elapsed time between the training and testing data periods.

1 Introduction

The capitalization task, also known as truecasing (Lita et al., 2003), consists of rewriting each word of an input text with its proper case information. The capitalization of a word sometimes depends on its current context, and the intelligibility of texts is strongly influenced by this information. Different practical applications benefit from automatic capitalization as a preprocessing step: when applied to speech recognition output, which usually consists of raw text, automatic capitalization provides relevant information for automatic content extraction, named entity recognition, and machine translation; many computer applications, such as word processing and e-mail clients, perform automatic capitalization along with spell corrections and grammar check.

The capitalization problem can be seen as a sequence tagging problem (Chelba and Acero, 2004; Lita et al., 2003; Kim and Woodland, 2004), where each lower-case word is associated to a tag that describes its capitalization form. (Chelba and Acero, 2004) study the impact of using increasing amounts of training data as well as a small amount of adaptation. This work uses a Maximum Entropy Markov Model (MEMM) based approach, which allows to combine different features. A large written newspaper corpora is used for training and the test data consists of Broadcast News (BN) data. (Lita et al., 2003) builds a trigram language model (LM) with pairs (word, tag), estimated from a corpus with case information, and then uses dynamic programming to disambiguate over all possible tag assignments on a sentence. Other related work includes a bilingual capitalization model for capitalizing machine translation (MT) outputs, using conditional random fields (CRFs) reported by (Wang et al., 2006). This work exploits case information both from source and target sentences of the MT system, producing better performance than a baseline capitalizer using a trigram language model. A preparatory study on the capitalization of Portuguese BN has been performed by (Batista et al., 2007).

One important aspect related with capitalization concerns the language dynamics: new words are introduced everyday in our vocabularies and the usage of some other words decays with time. Concerning this subject, (Mota, 2008) shows that, as the time gap between training and test data increases, the performance of a named tagger based on co-training (Collins and Singer, 1999) decreases.

This paper studies and evaluates the effects of language dynamics in the capitalization of newspaper

corpora. Section 2 describes the corpus and presents a short analysis on the lexicon variation. Section 3 presents experiments concerning the capitalization task, either using isolated training sets or by retraining with different training sets. Section 4 concludes and presents future plans.

2 Newspaper Corpus

Experiments here described use the RecPub newspaper corpus, which consists of collected editions of the Portuguese "Público" newspaper. The corpus was collected from 1999 to 2004 and contains about 148 Million words. The corpus was split into 59 subsets of about 2.5 Million words each (between 9 to 11 per year). The last subset is only used for testing, nevertheless, most of the experiments here described use different training and test subsets for better understanding the time effects on capitalization. Each subset corresponds to about five weeks of data.

2.1 Data Analysis

The number of unique words in each subset is around 86K but only about 50K occur more than once. In order to assess the relation between the word usage and the time gap, we created a number of vocabularies with the 30K more frequent words appearing in each training set (roughly corresponds to a freq > 3). Then, the first and last corpora subsets were checked against each one of the vocabularies. Figure 1 shows the correspondent results, revealing that the number of OOVs (Out of Vocabulary Words) decreases as the time gap between the train and test periods gets smaller.

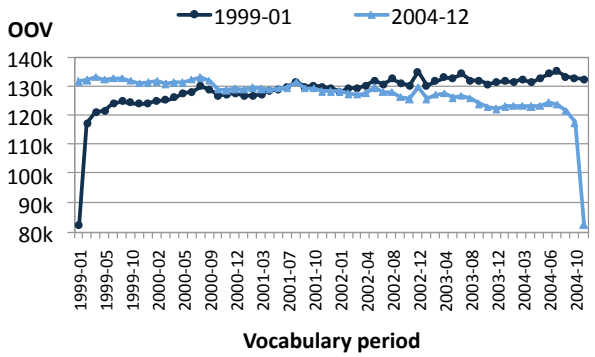

Figure 1: Number of OOVs using a 30K vocabulary.

3 Capitalization

The present study explores only three ways of writing a word: lower-case, all-upper, and first-capitalized, not covering mixed-case words such as "McLaren" and "SuSE". In fact, mixed-case words are also being treated by means of a small lexicon, but they are not evaluated in the scope of this paper.

The following experiments assume that the capitalization of the first word of each sentence is performed in a separated processing stage (after punctuation for instance), since its correct graphical form depends on its position in the sentence. Evaluation results may be influenced when taking such words into account (Kim and Woodland, 2004).

The evaluation is performed using the metrics: Precision, Recall and SER (Slot Error Rate) (Makhoul et al., 1999). Only capitalized words (not lowercase) are considered as slots and used by these metrics. For example: Precision is calculated by dividing the number of correct capitalized words by the number of capitalized words in the testing data.

The modeling approach here described is discriminative, and is based on maximum entropy (ME) models, firstly applied to natural language problems in (Berger et al., 1996). An ME model estimates the conditional probability of the events given the corresponding features. Therefore, all the information must be expressed in terms of features in a pre-processing step. Experiments here described only use features comprising word unigrams and bigrams: w_i (current word), $\langle w_{i-1}, w_i \rangle$ and $\langle w_i, w_{i+1} \rangle$ (bigrams). Only words occurring more than once were included for training, thus reducing the number of misspelled words. All the experiments used the `MegaM` tool (Daumé III, 2004), which uses conjugate gradient and a limited memory optimization of logistic regression. The following subsections describe the achieved results.

3.1 Isolated Training

In order to assess how time affects the capitalization performance, the first experiments consist of producing six isolated language models, one for each year of training data. For each year, the first 8 subsets were used for training and the last one was used for evaluation. Table 1 shows the corresponding capitalization results for the first and last testing sub-

Train	1999-12 test set			2004-12 test set		
	Prec	Rec	SER	Prec	Rec	SER
1999	**94%**	**81%**	**0.240**	92%	76%	0.296
2000	**94%**	**81%**	**0.242**	92%	77%	0.291
2001	94%	79%	0.262	93%	76%	0.291
2002	93%	79%	0.265	93%	78%	0.277
2003	94%	77%	0.276	93%	78%	0.273
2004	93%	77%	0.285	**93%**	**80%**	**0.264**

Table 1: Using 8 subsets of each year for training.

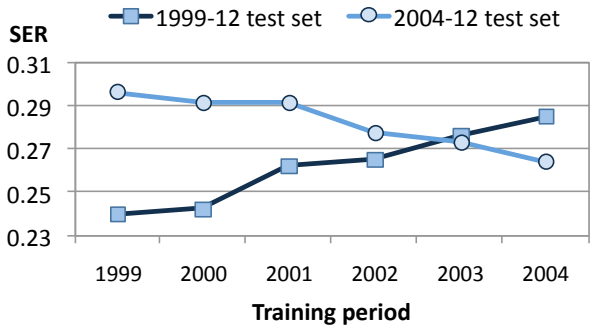

Figure 2: Performance for different training periods.

sets, revealing that performance is affected by the time lapse between the training and testing periods. The best results were always produced with nearby the testing data. A similar behavior was observed on the other four testing subsets, corresponding to the last subset of each year. Results also reveal a degradation of performance when the training data is from a time period after the evaluation data.

Results from previous experiment are still worse than results achieved by other work on the area (Batista et al., 2007) (about 94% precision and 88% recall), specially in terms of recall. This is caused by a low coverage of the training data, thus revealing that each training set (20 Million words) does not provide sufficient data for the capitalization task.

One important problem related with this discriminative approach concerns memory limitations. The memory required increases with the size of the corpus (number of observations), preventing the use of large corpora, such as RecPub for training, with

Evaluation Set	Prec	Rec	SER
2004-12 test set	93%	82%	0.233

Table 2: Training with all RecPub training data.

Checkpoint	LM #lines	Prec	Rec	SER
1999-12	1.27 Million	92%	77%	0.290
2000-12	1.86 Million	93%	79%	0.266
2001-12	2.36 Million	93%	80%	0.257
2002-12	2.78 Million	93%	81%	0.247
2003-12	3.10 Million	93%	82%	0.236
2004-08	3.36 Million	**93%**	**83%**	**0.225**

Table 3: Retraining from Jan. 1999 to Sep. 2004.

available computers. For example, four million events require about 8GB of RAM to process. This problem can be minimized using a modified training strategy, based on the fact that scaling the event by the number of occurrences is equivalent to multiple occurrences of that event. Accordingly to this, our strategy to use large training corpora consists of counting all n-gram occurrences in the training data and then use such counts to produce the corresponding input features. This strategy allows us to use much larger corpora and also to remove less frequent n-grams if desired. Table 2 shows the performance achieved by following this strategy with all the RecPub training data. Only word frequencies greater than 4 were considered, minimizing the effects of misspelled words and reducing memory limitations. Results reveal the expected increase of performance, specially in terms of recall. However, these results can not be directly compared with previous work on this subject, because of the different corpora used.

3.2 Retraining

Results presented so far use isolated training. A new approach is now proposed, which consists of training with new data, but starting with previously calculated models. In other words, previously trained models provide initialized models for the new train. As the training is still performed with the new data, the old models are iteratively adjusted to the new data. This approach is a very clean framework for language dynamics adaptation, offering a number of advantages: (1) new events are automatically considered in the new models; (2) with time, unused events slowly decrease in weight; (3) by sorting the trained models by their relevance, the amount of data used in next training stage can be limited without much impact in the results. Table 3 shows the re-

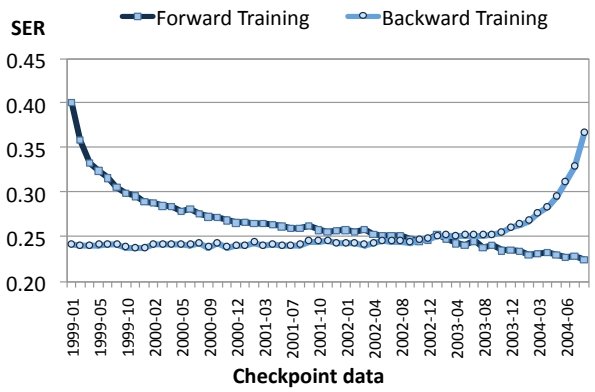

Figure 3: Training forward and backwards

sults achieved with this approach, revealing higher performance as more training data is available.

The next experiment shows that the training order is important. In fact, from previous results, the increase of performance may be related only with the number of events seen so far. For this reason, another experiment have been performed, using the same training data, but retraining backwards. Corresponding results are illustrated in Figure 3, revealing that: the backwards training results are worse than forward training results, and that backward training results do not allways increase, rather stabilize after a certain amount of data. Despite the fact that both training use all training data, in the case of forward training the time gap between the training and testing data gets smaller for each iteration, while in the backwards training is grows. From these results we can conclude that a strategy based on retraining is suitable for using large amounts of data and for language adaptation.

4 Conclusions and Future Work

This paper shows that maximum entropy models can be used to perform the capitalization task, specially when dealing with language dynamics. This approach provides a clean framework for learning with new data, while slowly discarding unused data. The performance achieved is almost as good as using generative approaches, found in related work. This approach also allows to combine different data sources and to explore different features. In terms of language changes, our proposal states that different capitalization models should be used for different time periods.

Future plans include the application of this work to BN data, automatically produced by our speech recognition system. In fact, subtitling of BN has led us into using a baseline vocabulary of 100K words combined with a daily modification of the vocabulary (Martins et al., 2007) and a re-estimation of the language model. This dynamic vocabulary provides an interesting scenario for our experiments.

Acknowledgments

This work was funded by PRIME National Project TECNOVOZ number 03/165, and FCT project CMU-PT/0005/2007.

References

F. Batista, N. J. Mamede, D. Caseiro, and I. Trancoso. 2007. A lightweight on-the-fly capitalization system for automatic speech recognition. In *Proc. of the RANLP 2007*, Borovets, Bulgaria, September.

A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

C. Chelba and A. Acero. 2004. Adaptation of maximum entropy capitalizer: Little data can help a lot. *EMNLP04*.

M. Collins and Y. Singer. 1999. Unsupervised models for named entity classification. In *Proc. of the Joint SIGDAT Conference on EMNLP*.

H. Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression.

J. Kim and P. C. Woodland. 2004. Automatic capitalisation generation for speech input. *Computer Speech & Language*, 18(1):67–90.

L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla. 2003. tRuEcasIng. In *Proc. of the 41st annual meeting on ACL*, pages 152–159, Morristown, NJ, USA.

J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. 1999. Performance measures for information extraction. In *Proceedings of the DARPA Broadcast News Workshop*, Herndon, VA, Feb.

C. Martins, A. Teixeira, and J. P. Neto. 2007. Dynamic language modeling for a daily broadcast news transcription system. In *ASRU 2007*, December.

Cristina Mota. 2008. *How to keep up with language dynamics? A case study on Named Entity Recognition*. Ph.D. thesis, IST / UTL.

Wei Wang, Kevin Knight, and Daniel Marcu. 2006. Capitalizing machine translation. In *HLT-NAACL*, pages 1–8, Morristown, NJ, USA. ACL.