The impact of Language Dynamics on the Capitalization of Broadcast News

Fernando Batista[1,2], *Nuno Mamede*[1,3], *Isabel Trancoso*[1,3]

[1]L2F – Laboratório de Sistemas de Língua Falada - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
[2]ISCTE – Instituto de Ciências do Trabalho e da Empresa, Portugal
[3]IST – Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal
{fmmb, njm, imt}@l2f.inesc-id.pt

Abstract

This paper investigates the impact of language dynamics on the capitalization of transcriptions of broadcast news. Most of the capitalization information is provided by a large newspaper corpus. Three different speech corpora subsets, from different time periods, are used for evaluation, assessing the importance of available training data in nearby time periods. Results are provided both for manual and automatic transcriptions, showing also the impact of the recognition errors in the capitalization task. Our approach is based on maximum entropy models, uses unlimited vocabulary, and is suitable for language adaptation. The language model for a given language period is produced by retraining a previous language model with data from that time period. The language model produced with this approach can be sorted and then pruned, in order to reduce computational resources, without much impact in the final results.

Index Terms: rich transcription, capitalization, discriminative methods, language dynamics, BN speech transcriptions.

1. Introduction

The capitalization, also known as truecasing [1], 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 the speech recognition output, which usually consists of raw text, it 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 [2, 1, 3], where each lower-case word is associated to a tag that describes its capitalization form. The impact of using increasing amounts of training data as well as a small amount of adaptation is studied by [2]. This work uses a Maximum Entropy Markov Model (MEMM) based approach, which allows the combination of different features. A large written newspaper corpora is used for training and the test data consists of BN data. The work of [1] describes a trigram language model (LM) with pairs (word, tag) estimated from a corpus with case information; the authors use 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 [4]. 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 LM. A previous study on the capitalization of Portuguese BN can be found in [5], where both generative and discriminative methods are used to perform capitalization of manual transcriptions.

New words are introduced everyday in the vocabularies and the usage of some other words decays with time. Concerning this subject, a study on Named Entity Recognition (NER) over written corpora was conducted by [6], showing that, as the time gap between training and test data increases, the performance of a named entity tagger based on co-training [7] decreases. This problem is also addressed in the work of [8], which proposes a daily adaptation of the vocabulary and LM to the topic of current news, based on texts daily available on the Web.

This paper addresses the capitalization task when performed over Broadcast News (BN) orthographic transcriptions. A written newspaper corpus provides the source of the capitalization information, and the evaluation is conducted on three different subsets of speech transcriptions, collected from different time periods. The importance of training data collected in nearby testing periods is also assessed, both for manual and automatic transcriptions. Only three ways of writing a word are explored: lower-case, all-upper, and first-capitalized, not covering mixed-case words such as "McLaren" and "SuSE". 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 capitalization of the first word of each sentence is assumed to be 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 [3], but the results described here do not consider them.

The paper is structured as follows: Section 2 summarizes the approach, which is based on maximum entropy models. Section 3 describes the properties of both written and spoken corpora. Section 4 investigates the capitalization when performed over manual orthographic transcriptions. Section 5 shows results for automatic transcriptions, and Section 6 concludes and presents future plans.

2. Capitalization method

In this study, we use a discriminative modeling approach, based on maximum entropy (ME) models, firstly applied to natural language problems by [9]. An ME model estimates the conditional probability of the events given the corresponding features. This framework provides a clean way of expressing and

Usage	Name	Recording period	Dur.	Words
Train		Oct. and Nov. 2000	61h	449k
Test	Eval	January 2001	6h	45k
	JEval	October 2001	13h	128k
	RTP07	May, June, Sep., Oct. 2007	6h	45k

Table 1: Different parts of the Broadcast News corpus.

combining several sources and different properties of events, such as word identification and POS (part-of-speech) tagging information. This approach requires all information to be expressed in terms of features, causing the resultant data file to become several times larger than the original one. This constitutes a training problem, making it difficult to train with large corpora. The classification however, is straightforward, making it interesting for on-the-fly usage.

The memory required for this approach increases with the size of the corpus (number of observations), preventing or making it difficult to use large corpora for training. For example, training with 4 million events requires about 8GB of RAM to process. This problem is mitigated by splitting the corpus into several subsets, and then iteratively retraining with each one separately. The first subset is used for training the first LM, which is then used to provide initial models for the next iteration over the next subset. This process goes on until all subsets are used. Although the final LM contains information from all corpora subsets, events occurring in the latest training sets gain more importance in the final LM. As the training is performed with the new data, the old models are iteratively adjusted to the new data. This approach provides a 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.

These experiments use only features comprising word identification, sometimes combined as bigrams: w_i (current word); $\langle w_{i-1}, w_i \rangle$, $\langle w_i, w_{i+1} \rangle$ (bigrams). All the experiments use the `MegaM` tool [10], which uses conjugate gradient and a limited memory optimization for maximum entropy classifiers.

3. Corpora description

3.1. Written newspaper corpus

Experiments described in this paper 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. It was split into subsets of about 2.5 Million words each, resulting in 59 subsets (between 9 to 11 per year). The last subset is used for evaluation. The original text was normalized and all the punctuation marks were removed, making it close to speech transcriptions, but without recognition errors. Only events occurring more than once were included for training, thus reducing the number of misspelled words and memory limitations.

3.2. Broadcast news corpus

The experiments described here also use an European Portuguese broadcast news corpus, originally collected for training and testing speech recognition and topic detection systems, in the scope of the ALERT European project [11]. The original corpus includes two different evaluation sets: Eval and JEval, the latter having been collected with the purpose of a "joint evaluation" set among all project partners. This corpus was complemented with a recent collection of 6 BN shows from the same public TV channel (RTP). Table 1 presents details for each part of the corpus. The manual orthographic transcription of this corpus constitutes the reference corpus, and includes information such as punctuation marks, capital letters and special marks for proper nouns, acronyms and abbreviations. Most of the corpus consists of planned speech. Nevertheless, 34% is still a large percentage of spontaneous speech.

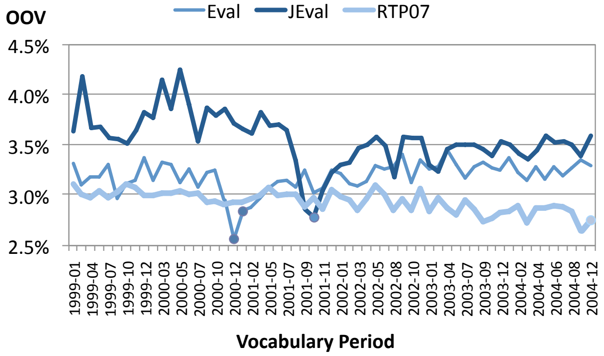

Figure 1: Proportion of words out of vocabulary.

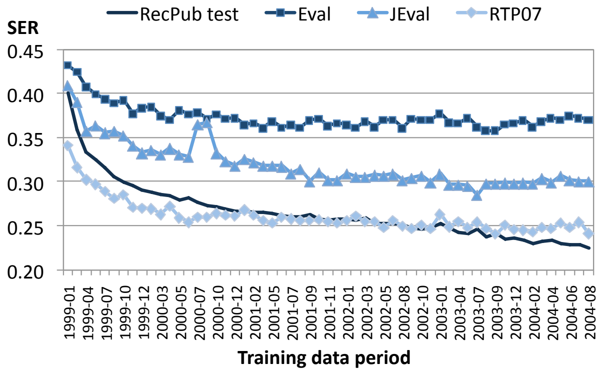

Figure 2: Capitalization evolution by training with RecPub.

Besides the manual orthographic transcriptions, we also have available the transcriptions produced by the automatic speech recognition (ASR) module, and other information automatically produced by the audio preprocessing (APP) module namely, the speaker id, gender and background speech conditions (clean/noise/music). Each word has a reference for its location in the audio signal, and includes a confidence score given by the ASR module.

4. Manual transcription results

Each subset of the newspaper corpus (about 2.5 million words) contains about 86K unique words, where only about 50K occur more than once. In order to assess the relation between the word usage and the language period, we created a vocabulary containing the 30K more frequent words appearing in each training corpus subset. Then we counted the number of words appearing in each testing set that were not covered by a given vocabulary. Figure 1 shows the correspondent results, where for each one of the testing periods the closest training periods are marked.

In order to assess the relation between the data period and

Training	Eval			JEval			RTP07		
Data	Prec	Rec	SER	Prec	Rec	SER	Prec	Rec	SER
1999	83%	80%	0.358	86%	84%	0.294	93%	80%	0.261
2000	83%	80%	0.359	86%	84%	0.289	92%	81%	0.257
2001	**84%**	**80%**	**0.345**	**87%**	**87%**	**0.264**	93%	80%	0.262
2002	84%	80%	0.355	86%	86%	0.275	93%	81%	0.251
2003	83%	79%	0.373	86%	85%	0.284	92%	82%	0.256
2004	84%	79%	0.361	87%	85%	0.283	**92%**	**82%**	**0.244**
All	83%	81%	0.352	**86%**	**88%**	**0.262**	**92%**	**85%**	**0.224**

Table 2: Retraining and evaluating with manual transcriptions.

Training	Eval			JEval			RTP07		
Data	Prec	Rec	SER	Prec	Rec	SER	Prec	Rec	SER
1999	72%	73%	0.552	75%	76%	0.494	79%	73%	0.460
2000	72%	73%	0.551	75%	77%	0.485	80%	73%	0.455
2001	**73%**	**73%**	**0.537**	**75%**	**78%**	**0.478**	80%	72%	0.461
2002	73%	73%	0.540	75%	78%	0.481	80%	73%	0.447
2003	72%	72%	0.558	74%	77%	0.493	79%	74%	0.455
2004	73%	73%	0.546	75%	77%	0.490	**80%**	**74%**	**0.443**
All	72%	74%	0.552	74%	80%	0.484	**79%**	**75%**	**0.443**

Table 4: Retraining with manual and evaluating with automatic transcriptions.

the capitalization performance, an experiment was conducted using all the newspaper training corpora and following the approach described in Section 2. Figure 2 shows the corresponding performance variations in terms of SER (Slot Error Rate) [12]. Only capitalized words (not lowercase) are considered as slots. Since retraining is used, the models calculated in a given period also include information from previous periods. The figure suggests that, for each test set, the performance grows until the correspondent time period is reached, but does not significantly improve after that period. The continuously growing performance for the RecPub and RTP07 test sets is related with their time period being later to the training data. The graph also shows that the performance is similar both to written corpora and to speech transcriptions, however, the performance evolution concerning the written corpora is smoother and steeper.

Another relevant question concerns the amount of training data required for achieving the best results. Is it necessary to use all the training data? Besides using all corpora for training, some experiments were also conducted using only the first 8 corpora subsets of every year for training (about 20 Million words). Previous experiments applying the written corpora LMs directly to the transcription data have revealed the problems of dealing with training and test sets with different properties. Hence, we have retrained the ME models calculated from written corpora with manual transcription training data, thus achieving 2% to 5% improved performance. Table 2 shows the corresponding results, in terms of SER, Precision and Recall, where all ME models were retrained with manual transcriptions after training with written corpora.

The first 6 rows corresponds to initial training with the first 8 corpora subsets of each year, while the last row corresponds to using all training data. The best results are shown on bold. The table shows that when all data is used, a better recall is achieved whereas the precision slightly decreases. In general we may conclude that, for capitalizing manual transcriptions, large amounts of training data are not necessary, if recent data is available. Results also show that the RTP07 test subset consistently presents best performances in opposition to the *Eval* subset. Nonetheless, the worse performance for the *Eval* and *JEval* sets is also due to the unusual topics covered in the news by that time (US presidentials and War on Terrorism).

5. Automatic transcription results

5.1. Alignment issues

Whereas the manual transcriptions already contain a reference capitalization, this is not the case of the automatic transcriptions. Therefore, in order to evaluate the capitalization task over this data, a reference capitalization must be provided. In order to do so, we have performed an alignment between the manual and automatic transcriptions, which is a non-trivial task mainly because of recognition errors. The alignment was performed using the NIST SCLite tool[1], followed by an automatic post-processing stage, either by correcting some SCLite basic errors or by aligning compound words which can be written/recognized differently. For example: the weekday "terça-feira" (Tuesday) sometimes is recognized as two isolated words "terça" (third) and "feira" (market). Table 3 presents statistics concerning the word alignment, where the post-processing corrections are shown in columns 2 (SCLite errors) and 3 (compounded words). Column 4 reveals that most of the words are correctly aligned, as expected. The WER (Word Error Rate) is shown in the last column, revealing the proportion of recognition errors in the corpus.

When in the presence of a correct word, the capitalization can be assigned directly, but insertions and deletions do not constitute a problem either. Moreover, most of the insertions and deletions consist of short functional words which usually appear in lowercase. Most of the alignment problems arise from the substitution errors where the reference word appears capitalized (not lowercase). In this case, three different situations may occur: (1) the two words have alternative graphical forms, a not infrequent phenomena in proper nouns, for example: ”Menezes” and "Meneses" (proper nouns); (2) the two words are different but share the same capitalization, for example: "Andreia" and "André"; and (3) the two words have different capitalization forms, for example "Silva" (proper noun) and "de" (of, from). We concluded, by observation, that most of the words in these conditions share the same capitalization if their lengths are similar. As a consequence, we decided to assign the same capitalization when the number of letters does not differ by more than 2 letters. The number of unsolved alignments (kept lowercase) are shown in the columns labeled "unsolved". From the table we can also see that there are more substitutions concerning first-capitalized words (e.g. proper nouns) than concerning uppercase words (e.g. acronyms).

5.2. Capitalization results

Table 4 shows the results of capitalizing automatic transcriptions with the LMs also used for table 2 results. As can be seen from the table, overall results are about 20% worse in terms of SER. Nevertheless, some errors may be due to unsolved alignment problems and more accurate results could be achieved if the capitalization alignment was manually corrected. These results also suggest a strong relation between the performance

[1] available from http://www.nist.gov/speech.

Corpus subset	Corrected align.		Correct	Del	Ins	Substitutions						WER
	sclite	comp. words				low	first is capitalized		all uppercase		other	
							solved	unsolved	solved	unsolved	unsolved	
Train	2138	282	420139	10193	25687	25841	2630	1721	637	112	87	14.5%
Eval	283	17	38172	1701	3122	5291	471	304	99	27	7	23.9%
JEval	781	98	103423	5647	6328	12745	1455	882	212	78	42	22.0%
RTP07	287	23	38986	1493	2776	4934	547	286	106	29	26	22.0%

Table 3: Alignment report, where: *corrected align* corresponds to SCLite errors corrected in a post-processing stage; *Correct* corresponds to correct alignments; *Del*, *Ins*, and *Substitutions* corresponds to the number word alignment deletions, insertions and substitutions; *low* corresponds to lower-case words; *Solved* and *unsolved* corresponds to alignments successfully or unsuccessfully (kept lowercase) performed.

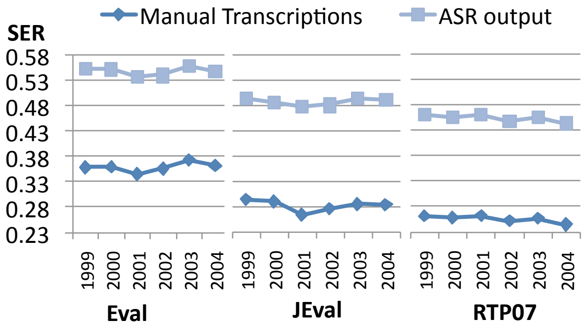

Figure 3: Comparing the capitalization results of manual and automatic transcriptions.

and the training period. The distribution of values is similar to the previous results concerning manual transcriptions. Figure 3 illustrates the relation between manual and automatic transcriptions, where the relation between training and testing periods becomes more clear. Some other tests were conducted, for example, by retraining with automatic transcriptions, but only small differences were achieved.

6. Conclusions and future work

This paper presented capitalization results for broadcast news transcriptions. The performance evolution is analyzed for three test subsets from different time periods. Capitalization results of manual and automatic transcriptions are presented, revealing the impact of the recognition errors on this task. For both types of transcription, the capitalization results show evidence that the performance is affected by the temporal distance between training and testing sets.

Together with a punctuation module, the capitalization module here described is applied to the subtitling stream automatically produced by the ASR module for the two daily BN shows of the public TV channel in Portugal, since early March. As explained, this module uses a dynamic lexical and LM adaptation from web materials. We are currently working on reusing these materials for retraining, taking advantage of the clean framework for language dynamics adaptation provided by maximum entropy models.

For the time being, only word identification features were used, but we are planning to use other alternative features, such as part-of-speech and clustering information. The word confidence score given by the recognition system was not used so far, but it will be included in future experiments.

7. Acknowledgments

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

8. References

[1] L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla, “tRuEcasIng,” in *Proc. of the 41st annual meeting on ACL*, (USA), pp. 152–159, ACL, 2003.

[2] C. Chelba and A. Acero, “Adaptation of maximum entropy capitalizer: Little data can help a lot,” *EMNLP ’04*, 2004.

[3] J.-H. Kim and P. C. Woodland, “Automatic capitalisation generation for speech input,” *Computer Speech & Language*, vol. 18, no. 1, pp. 67–90, 2004.

[4] W. Wang, K. Knight, and D. Marcu, “Capitalizing machine translation,” in *HLT-NAACL*, pp. 1–8, ACL, 2006.

[5] F. Batista, N. J. Mamede, D. Caseiro, and I. Trancoso, “A lightweight on-the-fly capitalization system for automatic speech recognition,” in *Proc. of the RANLP 2007*, (Borovets, Bulgaria), September 2007.

[6] C. Mota, *How to keep up with language dynamics? A case study on Named Entity Recognition*. PhD thesis, IST, Universidade Técnica de Lisboa (to be published), 2008.

[7] M. Collins and Y. Singer, “Unsupervised models for named entity classification,” in *Proc. of the Joint SIGDAT Conference on EMNLP*, 1999.

[8] C. Martins, A. Teixeira, and J. P. Neto, “Dynamic language modeling for a daily broadcast news transcription system,” in *Proc. of the ASRU 2007*, 2007.

[9] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, “A maximum entropy approach to natural language processing,” *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[10] H. Daumé III, “Notes on CG and LM-BFGS optimization of logistic regression.” http://hal3.name/megam/, 2004.

[11] H. Meinedo, D. Caseiro, J. P. Neto, and I. Trancoso, “Audimus.media: A broadcast news speech recognition system for the european portuguese language,” in *PROPOR’2003*, vol. 2721 of *LNCS*, pp. 9–17, Springer, 2003.

[12] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, “Performance measures for information extraction,” in *Proc. of the DARPA Broadcast News Workshop*, 1999.