Temporal Issues and Recognition Errors on the Capitalization of Speech Transcriptions*

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

[1] L^2F – Spoken Language Systems Laboratory - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://www.l2f.inesc-id.pt/
[2] ISCTE – Instituto de Ciências do Trabalho e da Empresa, Portugal
[3] IST – Instituto Superior Técnico, Portugal.

Abstract. This paper investigates the capitalization task over Broadcast News speech transcriptions. Most of the capitalization information is provided by two large newspaper corpora, and the spoken language model is produced by retraining the newspaper language models with spoken data. Three different corpora subsets from different time periods are used for evaluation, revealing 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 and uses unlimited vocabulary. 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.

Key words: capitalization, maximum entropy, discriminative methods, speech transcriptions, language dynamics.

1 Introduction

The capitalization task consists of rewriting each word of an input text with its proper case information. The intelligibility of texts is strongly influenced by this information, and different practical applications benefit from automatic capitalization as a preprocessing step. It can be applied to the speech recognition output, which usually consists of raw text, in order to provide relevant information for automatic content extraction, Named Entity Recognition (NER), and machine translation.

This paper addresses the capitalization task when performed over Broadcast News (BN) orthographic transcriptions. Written newspaper corpora are used as sources of capitalization information. The evaluation is conducted in three different subsets of speech transcriptions, collected from different time periods. The importance of training data collected in nearby testing periods is also evaluated.

* This work was funded by PRIME National Project TECNOVOZ number 03/165 and supported by ISCTE.

The paper is structured as follows: Section 2 presents an overview on the related work. Section 3 describes the approach. Section 4 provides the upper-bound results by performing the evaluation over written corpora. Section 5 shows results concerning speech transcriptions. Section 6 concludes and presents future plans.

2 Related Work

The capitalization problem can either be seen as a disambiguation problem or as a sequence tagging problem [1,2,3], 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 data is studied by [1]. 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 [2] describes 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 [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 language model. A previous study on the capitalization of Portuguese BN can be found in [5]. The paper makes use of generative and discriminative methods to perform capitalization of manual orthographic transcriptions.

The language dynamics is an important issue in different areas of Natural Language Processing (NLP): new words are introduced everyday and the usage of some other words decays with time. Concerning this subject, [6] conducted a study on NER over written corpora, showing that, as the time gap between training and test data increases, the performance of a named tagger based on co-training [7] decreases.

3 Approach Description

This paper assumes 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 [3]. Only three ways of writing a word will be considered here: lower-case, first-capitalized, and all-upper. Mixed-case words, such as “McLaren” and “SuSE”, are also treated by means of a small lexicon, but they are not evaluated in the scope of this paper.

The evaluation is performed using the metrics: Precision, Recall and SER (Slot Error Rate) [8]. Only capitalized words (not lowercase) are considered as slots and used by these metrics. For example: Precision is calculated by dividing the number of correctly capitalized words by the number of capitalized words in the test data.

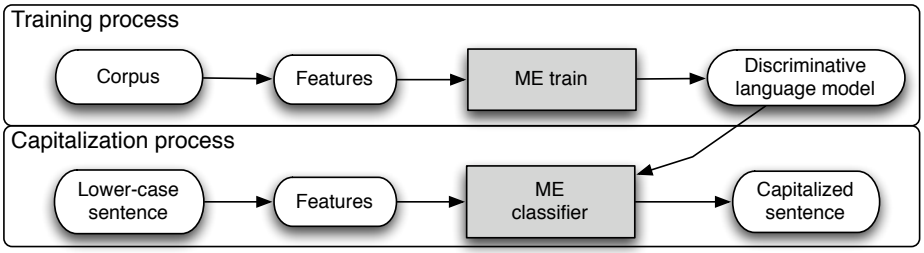

Fig. 1. Outline of the maximum entropy approach

3.1 The Method

The modeling approach used is discriminative, and is based on maximum entropy (ME) models, firstly applied to natural language problems in [9]. An ME model estimates the conditional probability of the events given the corresponding features. Figure 1 illustrates the ME approach for the capitalization task, where the top rectangle represents the training process using a predefined set of features, and the bottom rectangle illustrates the classification using previously trained models. This framework provides a very clean way of expressing and combining several knowledge sources and different properties of events, such as word identification and POS 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. 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 problem can be mitigated by splitting the corpus into several subsets. The first subset is used for training the first language model (LM), which is then used to provide initialized models for the next iteration over the next subset. This goes on until all subsets are used. The final LM contains information from all corpora subsets, but, 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; (3) by sorting the trained models by their relevance, the amount of data used in next training stage can be limited without much impact on the results.

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

Table 1. Newspaper corpora properties

Corpus	Usage	Period	#words
RecPub	train	September 1995 to June 2001	113.6 M
	test	2^{nd} Semester 2001	16.4 M
RecMisc	train	March to December 2007	19.2 M
	test	January 2008	1.3 M

Table 2. Forward and Backward training using unigrams and bigram features

	Training			LM	RecPub test			RecMisc test		
Exp	Corpus	Type	Last month	#Lines	Prec	Rec	SER	Prec	Rec	SER
1	RecPub	Back	1995-09	10.6 Million	92%	81%	0.258	93%	80%	0.250
2		Forw	2001-06	10.8 Million	**94%**	**82%**	**0.229**	94%	80%	0.238
3	RecMisc	Back	2007-03	5.2 Million	89%	75%	0.342	94%	85%	0.205
4		Forw	2007-12	5.2 Million	89%	75%	0.344	93%	85%	0.201
5	All	Back	1995-09	12.7 Million	91%	82%	0.256	92%	83%	0.228
6		Forw	2007-12	12.9 Million	90%	82%	0.268	**93%**	**87%**	**0.186**

4 Upper-bound performance using written corpora

This section presents results achieved for written corpora. Two different newspaper corpora are used, collected in separate time periods. The oldest and largest corpus is named RecPub and consists of collected editions of the Portuguese "Público" newspaper. RecMisc is a recent corpus and combines information from six different Portuguese newspapers, found on the web. Table 1 shows corpora properties and the corresponding training and testing subsets.

All the punctuation marks were removed from the texts, making them close to speech transcriptions, but without recognition errors. Only events occurring more than once were included for training, thus reducing the influence of misspelled words and memory limitations. The approach described in section 3 is followed, where the training corpus is split into groups containing a month of data. Table 2 shows the corresponding results. Each pair of lines in the table corresponds to using a given corpus, either by performing a normal training or training backwards. For example, both experiments 1 and 2 use RecPub training corpus, but while the training process of experiment 1 started at 2001-06 and finished at 1995-09, experiment 2 started at 1995-09 and finished at 2001-06. These two experiments use the same training data, but with a different training order. Results reveal that higher performances are achieved when the temporal difference between the time period of the last portion of training data and the time period of the testing data is smaller. The last month used in the training process seems to establish the period for which the LM is more adequate. The SER achieved in experiment 2 is better for both testing sets, given that their time period is closer to 2001-06 than to 1995-09. Notice, however, that experiments using different training sets can not be directly compared.

Table 3. Different parts of the Speech Recognition (SR) corpus

Sub-corpus	Recording period	Duration	Words
Train	2000 - October and November	61h	449k
Eval	2001 - January	6h	45k
JEval	2001 - October	13h	128k
RTP07	2007 - May, June, September, October	6h	45k

Table 4. Alignment report, where: *Cor, Ins*, *Del*, and *Sub* corresponds to the proportion of correct, insertions, deletions, and substitutions in terms of word alignments

Corpus	Alignment errors		Cor	Ins	Del	Sub				WER
part	c. words	sclite				lower	firstcap	allcaps	fail	
Train	282	2138	419872	25687	10193	25841	2630	637	1920	14.5%
Eval	17	283	38162	3122	1701	5291	471	99	338	23.9%
JEval	98	781	103365	6328	5647	12745	1455	212	1002	22.0%
RTP07	23	287	38983	2776	1493	4934	547	106	341	22.0%

5 Speech transcription results

The following experiments use the Speech Recognition corpus (SR) – an European Portuguese broadcast news corpus – collected in the scope of the ALERT European project [11]. Table 3 presents details for each part of the corpus. The original corpus included two different evaluation sets (Eval and JEval), and it was recently complemented with a collection of six BN shows, from the same public broadcaster (RTP07).

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. Each file in the corpus is divided into segments, with information about their start and end locations in the signal file, speaker id, speaker gender, and focus conditions. Most of the corpus consists of planned speech. Nevertheless, 34% is still a large percentage of spontaneous speech.

Besides the manual orthographic transcription, we also have available the automatic transcription produced by the Automatic Speech Recognition (ASR) module, and other information automatically produced by the Audio Preprocessor (APP) module namely, the speaker id, gender and background speech conditions (Noise/Clean). Each word has a reference for its location in the audio signal, and includes a confidence score given by the ASR module.

5.1 Corpus alignment

Whereas the reference capitalization already exists in the manual transcriptions, 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

Table 5. Retraining and evaluating with manual transcriptions

Training			Eval			JEval			RTP07		
Corpus	**Type**	**Last month**	**Prec**	**Rec**	**SER**	**Prec**	**Rec**	**SER**	**Prec**	**Rec**	**SER**
RecPub	Back	1995-09	84%	81%	0.347	86%	85%	0.287	92%	83%	0.243
	Forw	2001-06	**83%**	**81%**	**0.347**	**87%**	**86%**	**0.273**	**93%**	**83%**	**0.234**
RecMisc	Back	2007-03	82%	78%	0.388	85%	84%	0.312	91%	86%	0.217
	Forw	2007-12	81%	78%	0.403	84%	84%	0.313	91%	87%	0.215
All	Forw	2007-12	82%	80%	0.377	84%	87%	0.289	**91%**	**88%**	**0.206**

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 the recognition errors. Table 4 presents some issues concerning the word alignment. The alignment was performed using the NIST SCLite tool[4], but it was further improved in a post-processing step, either by aligning words which can be written differently or by correcting some SCLite basic errors. For example: the word "primeiro-ministro" (head of government) is sometimes written and recognized as two isolated words "primeiro" (first) and "ministro" (minister). The second and third columns present the number of corrected alignment errors.

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 functional words which usually appear in lowercase. The problem comes from the substitutions where the reference word appears capitalized (not lowercase). In this case, three different situations may occur: (1) the two words have different graphical forms, for example: "Menezes" and "Meneses" (proper nouns); (2) the two words are different but share the same capitalization, for example: "Andreia" and "André" (proper nouns); 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 do not differ by more than 2 letters. The column "fail" shows the number of unsolved alignments (kept lowercase).

5.2 Results over manual transcriptions

The initial capitalization experiments with speech transcriptions were performed with the LMs also used for table 2 results. Nevertheless, subsequent experiments have shown that the overall performance can be increased by retraining such models with speech transcription training data. By doing so, the SER performance increased about 3% to 5%. Table 5 shows the results concerning manual transcriptions, after retraining also with manual transcriptions. As expected, an overall lower performance is achieved when compared to written corpora,

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

Table 6. Retraining with manual and evaluating with automatic transcriptions

Training			Eval			JEval			RTP07		
Corpus	**Type**	**Last month**	**Prec**	**Rec**	**SER**	**Prec**	**Rec**	**SER**	**Prec**	**Rec**	**SER**
RecPub	Back	1995-09	72%	74%	0.546	74%	77%	0.502	79%	74%	0.459
	Forw	2001-06	**72%**	**74%**	**0.544**	**74%**	**78%**	**0.490**	**79%**	**74%**	**0.451**
RecMisc	Back	2007-03	72%	73%	0.558	73%	76%	0.516	79%	76%	0.441
	Forw	2007-12	71%	72%	0.579	73%	76%	0.517	79%	76%	0.445
All	Forw	2007-12	70%	73%	0.581	72%	79%	0.512	77%	77%	0.453

nonetheless, only a marginal difference is obtained for *RTP07*. The biggest difference is observed for the *Eval* and *JEval* test sets, however, *JEval* may be more representative, given that its size is almost three times the size of *Eval*. The smaller size of the RecMisc training data justifies the lower results achieved. The last two lines show the results when all the training is used. The relation between temporal issues and the performance can still be observed speech transcriptions but the differences are now much smaller, in part because manual transcriptions were used for retraining the final discriminative language models.

5.3 Results over automatic transcriptions

Table 6 shows the results of capitalizing automatic transcriptions. These experiments share the LMs also used for table 5 results. Other tests were conducted, for example, by retraining with automatic transcriptions, but only small differences were achieved. The overall SER decreased about 20%, however, these results are influenced by alignment problems and more accurate results can be achieved by manually correcting this capitalization alignment.

Figure 2 illustrates the differences between manual and automatic transcriptions. Results show that the RTP07 test subset consistently presents best performances in opposition to the *Eval* subset. The worse performance of *Eval* and

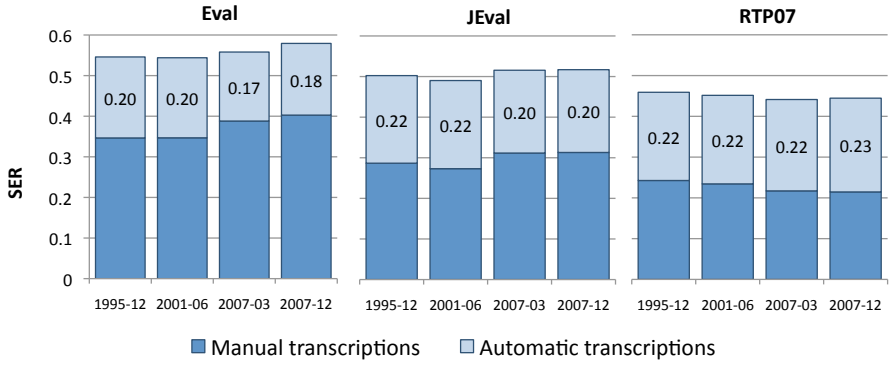

Fig. 2. Comparing the capitalization results of manual and automatic transcriptions

JEval is closely related with the main topics covered in the news by the time the data was collected (US presidentials and War on Terrorism). Results of the capitalization performed with RecPub for manual transcriptions suggest a relation between the performance and the training direction, and this relation can be found the same way in the speech transcriptions.

6 Conclusions and Future work

This paper have presented capitalization results, both on written newspaper corpora and broadcast news speech corpora. Capitalization results of manual and automatic transcriptions are compared, revealing the impact of the recognition errors on this task. Results show evidence that the performance is affected by the temporal distance between training and testing sets. Our approach is based on maximum entropy models, which provide a clean framework for language dynamics adaptation.

The use of generative methods in the capitalization of newspaper corpora is reported by [5]. Using WFSTs (Weighted Finite State Transducers) and a bigram language model, the paper reports about 94% precision, 88% recall and 0.176 SER. Using similar conditions, our approach achieves only about 94% precision, 82% recall and 0.229 SER. In the near future other features will be explored for improving the discriminative approach results. For example, the word confidence score given by the recognition system will be used in future experiments.

References

1. Chelba, C., Acero, A.: Adaptation of maximum entropy capitalizer: Little data can help a lot. EMNLP04 (2004)
2. Lita, L.V., Ittycheriah, A., Roukos, S., Kambhatla, N.: tRuEcasIng. In: Proc. of the 41^{st} annual meeting on ACL, Morristown, NJ, USA (2003) 152–159
3. Kim, J., Woodland, P.C.: Automatic capitalisation generation for speech input. Computer Speech & Language **18**(1) (2004) 67–90
4. Wang, W., Knight, K., Marcu, D.: Capitalizing machine translation. In: HLT-NAACL, Morristown, NJ, USA, ACL (2006) 1–8
5. Batista, F., Mamede, N., Caseiro, D., Trancoso, I.: A lightweight on-the-fly capitalization system for automatic speech recognition. In: Proc. of RANLP'07. (2007)
6. Mota, C.: How to keep up with language dynamics? A case study on Named Entity Recognition. PhD thesis, IST / UTL (2008)
7. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Proc. of the Joint SIGDAT Conference on EMNLP. (1999)
8. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proc. of the DARPA BN Workshop. (1999)
9. Berger, A.L., Pietra, S.A.D., Pietra, V.J.D.: A maximum entropy approach to natural language processing. Computational Linguistics **22**(1) (1996) 39–71
10. Daumé III, H.: Notes on CG and LM-BFGS optimization of logistic regression. (2004)
11. Meinedo, H., Caseiro, D., Neto, J.P., Trancoso, I.: Audimus.media: A broadcast news speech recognition system for the european portuguese language. In: PROPOR 2003. Volume 2721 of LNCS., Springer (2003) 9–17