TEXT NORMALIZATION AND SPEECH RECOGNITION IN FRENCH

Gilles Adda, Martine Adda-Decker, Jean-Luc Gauvain, Lori Lamel

Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{gadda,madda,gauvain,lamel}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this paper we present a quantitative investigation into the impact of text normalization on lexica and language models for speech recognition in French. The text normalization process defines what is considered to be a word by the recognition system. Depending on this definition we can measure different lexical coverages and language model perplexities, both of which are closely related to the speech recognition accuracies obtained on read newspaper texts. Different text normalizations of up to 185M words of newspaper texts are presented along with corresponding lexical coverage and perplexity measures. Some normalizations were found to be necessary to achieve good lexical coverage, while others were more or less equivalent in this regard. The choice of normalization to create language models for use in the recognition experiments with read newspaper texts was based on these findings. Our best system configuration obtained a 11.2% word error rate in the AUPELF 'French-speaking' speech recognizer evaluation test held in February 1997.

1. INTRODUCTION

The design of lexica and language models (LMs) are acknowledged to be important steps in the development of a speech recognizer and entail making some linguistic choices that are most often made without quantitative justification. In general these choices depend on the application and the language under consideration: much of recent speech recognition research for American English has been supported by ARPA and has been based on text materials which were processed to remove case distinction and compound words [11]. Case is generally kept as a distinctive feature in French and more importantly in German [8, 14].

The large amounts of training texts required for lexical and statistical language model design need to be cleaned and normalized before use. We compare different types of normalization of a source text containing 185 million words of the French newspaper *Le Monde*. The lexical coverages and language model perplexities for each text version were measured on a development text, with a lexicon containing the 64k most frequent words in the corresponding normalized training data. Our study shows which types of processing are most useful for maximizing the lexical coverage and what effect the processing has on the language model perplexity. The importance of lexical coverage and language model perplexity is illustrated by recognition experiments carried out in preparation for this year's AUPELF assessment of French recognizers.

2. NORMALIZATION OF FRENCH TEXTS

French is a language with high lexical variability stemming mainly from gender and number agreement (nouns, adjectives ...), and from verb conjugation. A given root form can have a large number of derived forms resulting in both low lexical coverage and poor language model training. The French language also makes frequent use of diacritic symbols which are particularly prone to spelling, encoding and formating errors.

Some of the normalization steps can be considered as baseline, such as the coding of accents and other diacritic signs (in ISO-Latin1), separation into articles, paragraphs and sentences, preprocessing of digits (10 000 → 10000), units (kg/cm^3), as well as the correction of typical newspaper formating and punctuation errors, and processing of unambiguous punctuation markers. Other kinds of normalization are generally carried out, but to the best of our knowledge, have never been systematically evaluated for speech recognizer development. We mention here:

N_0: processing of ambiguous punctuation marks (hyphen **-**, apostrophe **'**) not including compounds
N_1: processing of capitalized sentence starts
N_2: digit processing (110 → cent dix)
N_3: acronym processing (ABCD → A. B. C. D.)
N_4: emphatic capital processing (Etat → état)
N_5: decompounding (arc-en-ciel → arc en ciel)
N_6: no case distinction (Paris → paris)
N_7: no diacritics (énervé → enerve)

These elementary operations can be combined to produce different versions of normalized texts. Eleven such combinations are given in Table 1 using the normalizations listed above. Only the baseline normalizations are used to produce the reference text V_0. We use two large French dictionnaries: BDLEX [12] and DELAF [13] to produce V_1 and V_2 texts. A more detailed description of the normalizations can be found in [1].

While any normalization results in a reduction of information, the amount of information loss varies for the different types of normalizations. It is straightforward to recover a V_0 text (or an equivalent form) from a V_5 or V_6 text using some simple heuristics. For V_7 through V_{10} texts the original V_0 forms are nearly impossible to recover without additional knowledge sources. Furthermore V_9 and V_{10} texts seem poorly suited for speech recognition, as they produce high lexical ambiguity.

2.1 Lexical coverage

Recognizer vocabularies (word lists) are generally defined as the N most frequent words in training texts. We investigated the impact of different size training corpora using

	N_0	N_1	N_2	N_3	N_4	N_5	N_6	N_7	*Comment*
V_0	0	0	0	0	0	0	0	0	baseline normalizations
V_1	1	0	0	0	0	0	0	0	V_0 + ambiguous punctuations
V_2	1	1	0	0	0	0	0	0	V_1 + capitalized sentence starts
V_3	1	1	1	0	0	0	0	0	V_2 + digits
V_4	1	1	1	1	0	0	0	0	V_3 + acronyms
V_5	1	1	1	1	1	0	0	0	V_4 + emphatic capitalization
V_6	1	1	1	1	1	1	0	0	V_5 + decompounding
V_7	1	1	1	1	1	0	1	0	V_5 + case-insensitive
V_8	1	1	1	1	1	1	1	0	V_6 + case-insensitive
V_9	1	1	1	1	1	0	1	1	V_7 + no diacritics
V_{10}	1	1	1	1	1	1	1	1	V_8 + no diacritics

Table 1: For each version V_i ($i = 0, \ldots, 10$) of normalized text, the elementary normalization steps N_j ($j = 0, \ldots, 7$) are indicated by 1 in the corresponding column.

version specific 64k word lists from the *Le Monde* training texts. The three training text sets compared are:

T_0 : years 1987-88 (40M words)[1]
T_1 : years 1987-95 (185M words)
T_2 : years 1991-95 (105M words)[2]

In order to measure lexical coverage of the different normalized text versions, we selected a development text set containing about 20000 words from the *Le Monde* newspaper taken from the month of May 1996[3] (**dev0**).

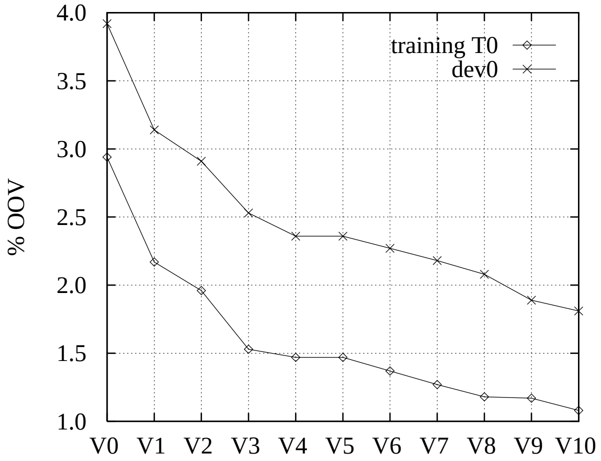

Figure 1: OOV rates for different normalization versions V_i on T_0 training data and dev0 test data using 64k word lists.

The out of vocabulary (OOV) word rate for the 11 versions of normalized T_0 training texts is shown in Figure 1 as measured on the **dev0** test text for a 64k word list. The evolution of the OOV rate, as a function of the normalized text versions, is observed to be the same for both the training and the development test data. A large reduction in OOV rate is obtained for the V_1, V_2 and V_3 text versions, which correspond to the processing of ambiguous punctuation marks, sentence-initial capitalization, and digits. Subsequent normalizations improve coverage, but to a lesser extent.

A difference in lexical coverage of about 20% is observed between T_0 training and dev0 test data. While this could be due to the relatively small size of the T_0 training text, the main cause is the large time gap of (about 8 years) between the text sets. Using larger and more recent training texts (T_1 or T_2) this difference can be reduced to 1%. We have noticed that the time proximity between training and test data is more important than the use of additional, but older data in minimizing the OOV rate. This is shown in Figure 2, where equivalent OOV rates on the dev0 test data were obtained for T_2 (105 M words) and the T_1 (185 M words) training data. Thus, the selection of training data for a given test condition is seen to be more important than the effect of many of the elementary normalization steps (compounding, case-sensitivity).

Optimized training data selection is carried out by weighting recent training texts more than the older text material. This optimization can even erradicate the effects of some minor normalizations. The optimized word list used in the recognition experiments in Section 3 has the same number of OOV words for both the V_5 and V_6 text versions.

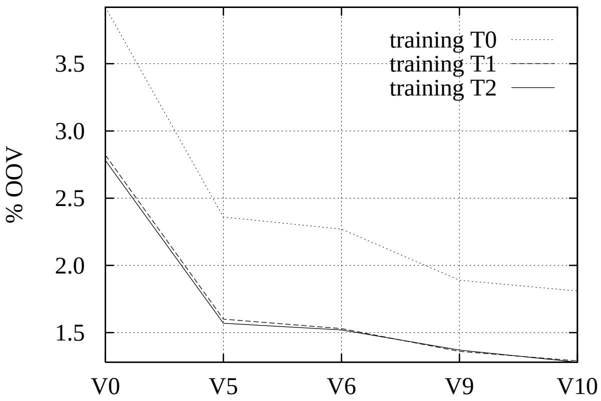

Figure 2: OOV rates for normalization versions V_0, V_5 V_6, V_9 and V_{10} on dev0 test data using 64k word lists derived from different training text sets: T_0 (40M words), T_1 (185M words) and T_2 (105M words).

2.2 Language model perplexity

The characteristics of each text version V_i ($i = 0, \ldots, 10$), in terms of the total number of words and number of distinct words, directly influences the language model properties. Normalizations of type N_0, N_1, N_2 considerably reduce the number of different word forms, while increasing the total number of words in the corpus as shown in Figure 3. This should be in favor of more reliable LM training. However, better language model accuracy may be achieved if a larger number of different word forms are considered, provided that they are linguistically meaning-

[1] These were baseline resources for all partners in the AUPELF French recognizer evaluation project.

[2] T_2 is significantly smaller than T_1, but contains on average more recent data.

[3] This corresponds to the time period from which the AUPELF development test data (**dev**) were selected.

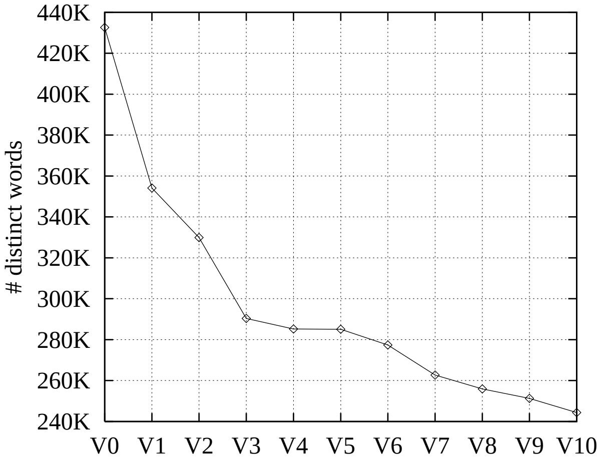

Figure 3: Number of different words in the training text T_0 for different normalization combinations V_i.

ful and there are sufficient training data available.

Perplexity is commonly used to measure LM efficiency, and a decrease in perplexity generally entails a decrease in error rate. Precise perplexity comparisons make sense only if the perplexities have been estimated on identical test texts. In order to compare LM perplexities of the different normalized versions V_i of the test set W with different text lengths[4] we use a normalized perplexity measure, where the standard perplexity measure $p = \Pr(\text{W}|\text{LM})^{-\frac{1}{n}}$ is transformed as follows [6]:

$$p^* = p^{\frac{n_i}{n_r}}$$

where n_i is the length of W_i (the normalized version V_i of W) and n_r is the length of a reference version W_r.

The basic idea here is to consider that $\Pr(\text{W}_i|\text{LM}_i)$ can be compared for different normalizations W_i of W. OOV words in the text are replaced by the symbol <UNKNOWN>, and ignored when computing the perplexity. Figure 4 shows the normalized and unnormalized perplexity values of the dev0 text set, using a T_0 64k trigram LM. The normalized perplexity grows with increasing normalizations, but the largest effect is seen for the V_1 text (ambiguous punctuations) and the V_4 text (acronyms). Perplexity measures for V_5 and V_6 text versions of a T_1 64k trigram LM are also shown. The increase in training data yields a relative perplexity reduction of about 10%.

Our investigations with different text normalizations indicate that the precise choice of normalization is unlikely to be crucial for the speech recognizer. Thus, for recognition experiments we chose the V_5 normalization, which seems to be a good tradeoff between the best possible coverage and perplexity values, and providing as output a reasonably correct form of written French.

3. SPEECH RECOGNITION

Previous experiments in large vocabulary speech recognition in French have been reported in [8] using a 20k vocabulary (Esprit-SQALE project) on test sets with a controlled OOV rate of about 2%. Without artificial limitation the OOV rate tends to be closer to 5 or 6%. Hence there is a need for larger vocabularies in French, which in turn require larger text corpora for language model training.

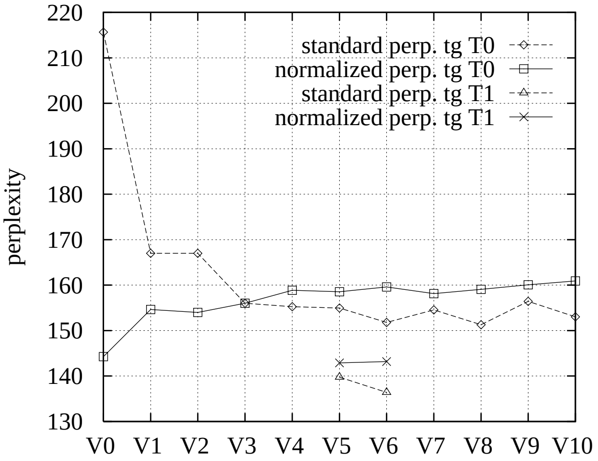

Figure 4: Perplexities of dev0 text (standard & normalized) for different normalization versions V_i using 64k trigram LMs estimated from T_0 (40 Mwords) and T_1 (185 Mwords) training texts.

The recognition system configuration is extensively described in [2]. We summarize here the main characteristics concerning the results presented below.

Acoustic Modeling The acoustic parameters consist of 39 cepstral parameters (including first and second order derivatives) derived from a Mel spectrum estimated on a 8kHz bandwidth. Each acoustic model is a 3-state left-to-right CDHMM representing a phone in context. Gender-dependent models are used. The models were trained using 66,585 sentences from 120 speakers (BREF[9]).

Language Modeling We used 65k bigram and trigram LMs trained on 200M words of *Le Monde* and *Le Monde Diplomatique* texts (years 1987-1996), and 70M words from *Agence France Presse* (AFP, years 1994-1996, distributed by LDC).

Lexicon The training and recognition lexica were developed at LIMSI. Each lexical entry is phonemically transcribed using a 34 phone set including silence. A pronunciation graph is associated with each word in order to account for pronunciation variants.

Decoding Decoding is carried out in 3 passes: The first pass uses a bigram language model (2.2M bigrams) to generate a word graph. The acoustic models used in this pass consist of about 3000 position-dependent triphones with about 8000 tied states. The second decoding pass, makes use of the word graph with a trigram LM (14M bigrams and 22M trigrams), and position-independent triphone models (about 9000 tied states are distributed among over 5000 models). Prior to the third decoding pass unsupervised acoustic model adaptation based on MLLR [10] is carried out using the hypotheses generated in the second pass. An interpolated language model based on word trigrams and class bigrams [7] is optionally used in this pass.

These experiments were carried out within the AUPELF project using two test sets (dev-T, eval-T) each containing

[4] A 1% variation in text length yields a 5% variation of the perplexity if $p = 150$.

about 600 sentences (15000 words). For each set T, a subset T' of 300 sentences contains the paragraphs with the lowest OOV rates.

Table 2 shows the results of recognition experiments using a standard trigram and a trigram+biclass language model. The standard trigram uses a backoff procedure to word bigrams and unigrams, whereas the trigram+biclass model [7] interpolates trigrams with class bigrams. The interpolated model yields a small perplexity decrease from 135 to 131 on the dev-T text, and a small but consistent error reduction across both test sets.

test	*tg stand.*	*tg+biclass*
65k-dev	12.9	12.7
65k-eval	11.5	11.2

Table 2: Word error rates on the development and the evaluation sets, using a 65k vocabulary and trigram LM (standard trigram and biclass interpolation).

word list	*20k*	*30k*	*40k*	*50k*	*60k*	*65k*
T, std	6.38	4.30	3.15	2.36	1.95	1.79
T, opt	6.15	4.04	2.80	2.11	1.59	1.34
T', std	3.60	2.37	1.70	1.22	0.96	0.87
T', opt	3.51	2.03	1.35	0.97	0.58	0.44

Table 3: OOV rates on the dev T and T' sets, for word lists ranging from 10k to 65k words. The word lists consist of the N most frequent words in the T_0 baseline training data (std=standard) or optimized over the available training data (opt=optimized).

The impact of word list size and training data selection (optimization) on lexical coverage is shown in Table 3. Optimization yields an absolute gain of about 0.3% for dev-T and for dev-T' regardless of the word list size. Lexical coverage can thus be improved by increasing and optimizing the system's vocabulary.

As can be seen in Table 4 filtering the test data from T to T' not only reduces the OOV rate, but the perplexity drops significantly. The large word error reduction (4 times the OOV reduction) is explained by both OOV and perplexity decreases, with a major contribution due to the difference in perplexity. Using the same recognition system with the same LM on the complete test set T, but outputting only words of a 20k word list we can measure the impact of lexical coverage independently of other factors. By simulating an increase in coverage from 20k to 65k the observed error reduction is 60% of the OOV reduction. This illustrates that the probability of misrecognizing infrequent words is high, as all OOVs would be recovered by a perfect recognizer.

Comparing independent 20k and 65k systems we have observed a word error reduction of about 1.3 times the reduction in OOV. Based on these measures we infer that an OOV generates on average 2.2 errors.[5]

4. DISCUSSION

We have investigated different types of normalizations for French newspaper texts and measured their impact on lexical coverage and LM perplexity. These parameters are known to be related to speech recognition accuracy, as demonstrated by the recognition experiments.

test	*OOV (%)*	*ppx*	*Werr (%)*
65k-dev-T	1.34	135	12.9
65k-dev-T'	0.45	105	8.9

Table 4: OOV rates, perplexity and word error rates on the dev-T and dev-T' sets, using a 65k vocabulary and a trigram LM.

Normalizations resulting in significant reductions of the OOV rate (N_0-N_2) have been identified. A strong correlation of the text version (V_i) with both number of distinct word forms (Fig. 3) and with the normalization-dependent OOV rate (Fig. 2) was observed. Similar perplexity values were obtained for most of the normalizations explored. The largest changes in perplexity were observed after treatment of ambiguous punctuations and of acronyms.

Our investigations have shown that some normalizations should be systematically applied (typically $N_0 - N_2$). Different more complex combinations give approximately equivalent results in terms of coverage and perplexity. The final choice among these depends on the application.

REFERENCES

[1] G. Adda , M. Adda-Decker, “Normalisation de textes en français: une étude quantitative pour la reconnaissance de la parole”, *1ères JST FRANCIL*, Avignon, April 1997.

[2] G. Adda , M. Adda-Decker, J.L. Gauvain, L. Lamel, “Le système de dictée du LIMSI pour l'évaluation AUPELF'97”, *1ères JST FRANCIL*, Avignon, April 1997.

[3] M. Adda-Decker, G. Adda, L. Lamel, J.L. Gauvain, “Developments in Large Vocabulary, Continuous Speech Recognition of German,” *IEEE ICASSP-96*, Atlanta, 1996.

[4] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, “Speaker-independent continuous speech dictation,” Speech Communication **15**, pp. 21-37, Sept. 1994.

[5] J.L. Gauvain, L. Lamel, G. Adda, J. Mariani, “ Speech-to-Text Conversion in French,” *Int. Journal of Pattern Recognition and Artificial Intelligence*, **8**(1), Jan. 1994.

[6] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, “Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task,” *IEEE ICASSP-96*, Atlanta, 1996.

[7] M. Jardino, “Multilingual stochastic n-gram class language models,” *IEEE ICASSP-96*, Atlanta, 1996.

[8] L. Lamel, M. Adda-Decker, J.L. Gauvain, “Issues in Large Vocabulary, Multilingual Speech Recognition,” *Eurospeech'95*, Madrid, Sept. 1995.

[9] L.F. Lamel, J.L. Gauvain, M. Eskénazi, “BREF, a Large Vocabulary Spoken Corpus for French,” *EuroSpeech'91*, Genoa, Sept. 1991.

[10] C.J. Legetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech & Language*, **9**, pp. 171-185, 1995.

[11] D.B. Paul, J.M. Baker, “The Design for the Wall Street Journal-based CSR Corpus,” *ICSLP'92*, Banff, Oct. 1992.

[12] G. Pérennou, “Le projet BDLEX de base de données lexicales et phonologiques,” *1ères journées du GRECO-PRC CHM*, EC2 éd., Paris, 24-25 November 1988.

[13] M. Silberztein, *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, 1993.

[14] S.J. Young et al., “Multilingual large vocabulary speech recognition: the European SQALE project,” *Computer Speech & Language*, **11**(1), pp. 73-89, Jan. 1997.

[5] We estimate the average number of errors caused by an OOV to be the ratio of the difference in error rate between the 20k and 65k systems, and the rate of 20k-OOVs recovered.