Lexicon Optimization for Dutch Speech Recognition in Spoken Document Retrieval

Roeland Ordelman, Arjan van Hessen, Franciska de Jong

Department of Computer Science
University of Twente, The Netherlands
{ordelman,hessen,fdejong}@cs.utwente.nl

Abstract

In this paper, ongoing work concerning the language modelling and lexicon optimization of a Dutch speech recognition system for Spoken Document Retrieval is described: the collection and normalization of a training data set and the optimization of our recognition lexicon. Effects on lexical coverage of the amount of training data, of decompounding compound words and of different selection methods for proper names and acronyms are discussed.

1. Introduction

In a series of related collaborative projects in which among others both the University of Twente and the research organization TNO participate, a Multimedia Information Retrieval (MMIR) environment for Dutch video archives is being developed that incorporates Dutch speech recognition (SDR). In this paper lexicon optimization of Dutch speech recognition in a SDR task, is discussed as deployed in the projects DRUID (Document Retrieval Using Intelligent Disclosure) [2] and ECHO (European CHronicles Online) [3]. Whereas DRUID concentrates on contemporary data from Dutch broadcasts, the ECHO project aims mainly at the disclosure of *historical* national video archives.

Dutch text data that is suitable for language modelling is not available in as large amounts and in that variety as for English. Since for the transcription of broadcast news large vocabulary speech recognition is regarded as most appropriate (cf. the TREC tasks for Spoken Document Retrieval [8]), one needs large amounts of data to be able to estimate the large vocabulary language model parameters reliably. Also, the text data should be close to the task domain, which requires at least a careful source selection and preferably text corpora with (manually) tagged domains. With this in mind, the collection and preparation of text data set suitable for the envisaged tasks, the transcription of broadcast news and historical video archives, was started as described in section 2

In automatic speech recognition, the goal of lexicon optimization is to construct a lexicon with exactly those words that are most likely to appear in the test data. Lexical coverage of a lexicon should be as high as possible to minimize out-of-vocabulary (OOV) words. Firstly, lexical coverage of lexicons is compared given different amounts of training data. Next, we discuss lexical coverage across languages, addressing the problem of word compounding, an important characteristic of Dutch compared to other languages. Word compounding increases lexical variety of a language which complicates lexicon optimization. Decompounding compound words into to their separate constituents improved lexical coverage for German [5]. Results on different decompound procedures for Dutch are reported.

Another way to improve lexical coverage, especially in a SDR environment, is the accurate selection of proper names and acronyms. These are important, information-carrying words and an optimal recognition of these types of words is therefor crucial. Different selection methods are compared in section 3.3.

2. Text Data collection

From various sources, a total of 152M words of text were collected for language modelling. From the 'Persdatabank', an organization that administers the exploitation rights of four major Dutch newspapers, we received Dutch newspaper data (~147M words from 1994, 1995, 1999, 2000). In [4] language model (LM) perplexity on broadcast news test data is reduced considerably by adding transcripts of broadcast news shows (BNA & BNC corpus) to the LM training data. Since similar corpora are not available for Dutch, we started recording teletext subtitles from broadcast news and 'current affairs' shows in 1998. On top of that the Dutch National Broadcast Foundation (NOS) provides the auto cues of broadcast news shows. Although the teletext material, and in a lesser degree the auto cues material, do not match as good as manual transcripts, they are a welcome addition to our data set. The total amount of 152M words in our collection, is surely not comparable with the amounts of text data that is available for English for example, but at least provides a reasonable basis to start from.

2.1. Spelling Variants

To allow for content selection (foreign affairs, politics, business, sports, etc.) for the creation of domain specific language models, all data was first converted to XML and stored in a database. A pre-processing module was build on top of the database to enable the normalization of the raw newspaper text to a version more suitable for language modelling purposes. Basically, the module reduces the amount of spelling variants: it removes punctuation (or writes certain punctuation's to a special symbol), expands numbers, currencies and abbreviations, and does case processing based on the uppercase/lowercase statistics of all newspaper data in the corpus.

In the case processing step, all word frequencies were collected and grouped ignoring case distinction. For every group, the ratio (ρ) of the two highest word frequencies was computed. Every word in a single group was either written to the variant with the highest word frequency or was kept unchanged. The latter was done when its frequency exceeded a word frequency threshold (N) and $\rho < 0.7$. The ρ variable was used to rule out words that are very frequently written in the wrong case (like

'ROTTERDAM' appearing in newspaper headings) and to keep words/names that are also frequently occurring names/words like in 'minister president *Kok* (prime minister Kok)' and 'de *kok* bereidde een maaltijd (the cook made a meal)'. Optimal values of N (500) and ρ were determined empirically on some training data.

Normalization	#words	# distinct
None	20M	879K
Punctuation	22,4M (+11%)	387K (-56%)
Case processing	24,2M (+7%)	341K (-12%)
Spelling	24,2M	323K (-5%)

Table 1: Number of words and distinct words after different processing steps

Finally, the module tries to correct frequent spelling errors based on a (preferred) spelling suggestion list of ~800K words that was provided by the Dutch dictionary publisher Van Dale Lexicography. According to the spelling suggestion list (number of suggestions ranging from 1 to 5, ranked by the amount of effort needed (insertions, deletions, etc) to get to the correct form), about 27% of the words in our data set was either not written conform the preferred spelling or did not exist in the spelling checker dictionary. Uppercase words were excluded from spelling correction since the suggestions of these words appeared to be unreliable. The lowercase words were corrected according to the following procedure: given a incorrectly spelled word that has a frequency > 10, frequencies of occurrence in the newspaper data of the word itself and all suggestions were looked up. Also a ratio (υ) of the frequency of the incorrect word and the highest suggestion frequency was computed. Since we wanted to be minimize correction errors, the suggestion with the highest frequency won only if $\upsilon < 0.5$. In Table 1 the number of words and distinct words in a subset of the data, newspaper data from 2000 (further called PDB), are listed before and after the different normalization steps.

3. Lexicon optimization

3.1. Data selection

In [9] best coverage of the lexicon (measured in OOV rate on some test data) was investigated by accumulating available text data in portions of 5M words. More training data resulted in lower OOV rates, but improvements slowed down considerably after 30M-50M words. We repeated this experiment using all newspaper data (~147M words, data was added in decreased order of recency) and investigated OOV rates using lexicons of 20K, 40K and 60K words.

To obtain out-of-vocabulary statistics, we used a test set of 35K words of teletext subtitling information from January 2001 broadcast news (BN), as a rough estimate of actual transcriptions of broadcast news shows. In Figure 1 the OOV rates of a 20K, 40K and 60K word list are shown. Best results are listed in table 2. As expected OOV rates decreased with growing amounts of data although improvements slowed down at around 60M words instead of 30M-50M words as in [9]. This could be due to the poor lexical coverage of Dutch compared to English as discussed in section 3.2. Neither did OOV rates go up again after a minimum is reached. It could well be however, that OOV rates indeed get worse when even more data is added (up to 300M words as in [9]).

Note that there is a sudden gap in the range of 45M to 60M words. This is most probably caused by our unbalanced newspaper data set (in time) along with the fact that the data was added in decreased order of recency: the first 45M words are from the years 1994 and 1995; the words that are accumulated next come from the year 1999 and 2000. As described in [9], recency has a relatively strong effect on lexical coverage, which explains the sudden drop in OOV rate.

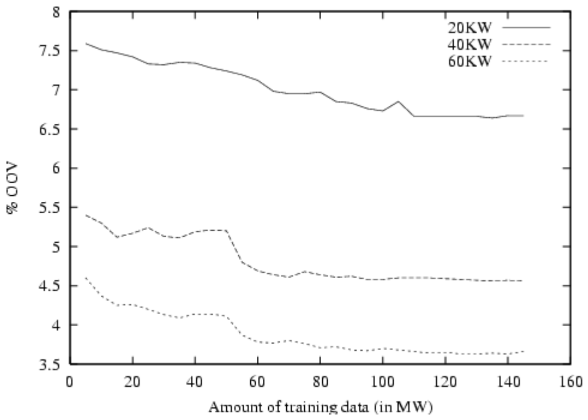

Figure 1: Effect on OOV rate

Lexicon size	Amount of data	OOV
20K	110M	6.66%
40K	145M	4.56%
60K	125M	3.63%

Table 2: Best OOV rates

3.2. Compounding

In [7] lexical variety and lexical coverage is compared across languages with the ratio:

$$\frac{\#words}{\#distinctwords}$$

In Table 3 the statistics found in [7] are given and those for Dutch are added (coverage based on the normalized training text). It shows that Dutch is comparable with German although lexical coverage of German is even poorer than lexical coverage of Dutch. The reason is that German has case declension for articles, adjectives and nouns, while Dutch has not. The major reason for the poor lexical coverage of German compared to the other languages is word compounding [5, 6]. Words can (almost) freely be joined together to form new words. In automatic speech recognition, the goal of lexicon optimization is to construct a lexicon with exactly those words that are most likely to appear in the test data. Lexical coverage of a lexicon should be as high as possible to minimize out-of-vocabulary (OOV) words. Because of compounding in German, a larger lexicon is needed to achieve the same lexical coverage as for English.

Since in Dutch compounding is frequently applied as well, we investigated the effect on lexical coverage and OOV rates by decompounding Dutch compound words into their separate

	Eng	It	Fr	Du	Ge
#words	37,2M	25,7M	37,7M	22M	36M
#distinct	165K	200K	280K	320K	650K
ratio	225	128	135	69	55
5K cov	90,6%	88,3%	85,2%	84,6%	82,9%
20K cov	97,5%	96,3%	94,7%	93%	90,0%
65K cov	99,6%	99,0%	98,3%	97,5%	95,1%
20K OOV	2,5%	3,7%	5,3%	7%	10,0%
65K OOV	0.4%	1.0%	1.7%	2,5%	4,9%

Table 3: Comparison of languages in terms of number of distinct words, lexical coverage and OOV rates for different lexicon sizes.

constituents. Decompounding however, is complicated and actually requires a refined morphological analysis. Since we do not have tools for such an analysis, we used the following decompounding procedure: every word is checked upon a 'dictionary' list of 217K frequent compound words (provided by Van Dale Lexicography). Every compound is translated into two separate constituents. After a first run, all words are checked again in a second run, to split compound words that remained because they originally consisted of more then two constituents.

In Table 4 the results of such a (full) decompounding procedure are shown: 66K distinct compound words were decompounded in the full decompounding case. However, the decrease in distinct words is only 62K, meaning that 4000 *new* words were created by the procedure. This happens when at least one of the constituents does not appear as a separate word (in the data set). When we look at lexical coverage results, it appears that full decompounding did not improve lexical coverage at all, in spite of the enhanced ratio statistic.

Language	Original	Decompounding	
		Full	Partial
# words	22,02M	22,14M	22,12M
#distinct words	321K	259K (-66K)	266K (-55K)
ratio	69	86	83
5K cov	84,6%	84,4 (-0,2)%	84,8 (+0,2)%
20K cov	93%	93% (0)	93,4 (+0,4)%
65K cov	97,5%	97,4 (-0,1)%	97,9 (+0,4)%

Table 4: Comparison of Dutch (original and decompounded)

Although for every decompounded word the ratio improves, this is not always the case for lexical coverage. When for example both a compound word as its constituents do not appear frequently enough to be selected in a lexicon, decompounding this compound will result in an increase in OOV rate: without decompouding the missing compound results in only one OOV word, after decompounding the two missing constituents result in two OOV words.

In an attempt to deal with this effect, a 'partial' decompounding procedure was devised, in which decompounding was applied only if both constituents of the compound already existed in the full word list of the training text. Using this procedure, 55K distinct compounds were decompounded, resulting in a 55K decrease of distinct words. Although the ratio statistic deteriorated a little, lexical coverage slightly improved. The absolute gain of approximately 0,4% with the 65K lexicon is not that high (0,7%) as reported by [5] however. In this study, an absolute gain of 0.7% was achieved after rule-based decompounding.

It could be argued that producing new or infrequent words by decompounding has a much smaller effect when a larger data set is used. Words that seem to be very infrequent (since they are non-existent) in a small data set could, after decompounding is applied, appear frequently enough in a larger data set to become significant. To investigate this, the data accumulation experiment described in section 3.1 was repeated, this time using data that was decompounded without any restrictions. In figure 2 the OOV rates of the various 60K lexicons are compared with and without decompounding the accumulated data. Although according to our expectations OOV rates decreased after full decompounding, the results did not show that lexical coverage hardly improves when only a small amount of data is used, as we have seen earlier: OOV rates are *consistently* smaller. This might well be due to the test set that is used. In the experiment last mentioned, OOV rates were obtained by testing on teletext subtitling data (BN), instead of the newspaper data as in the former experiment.

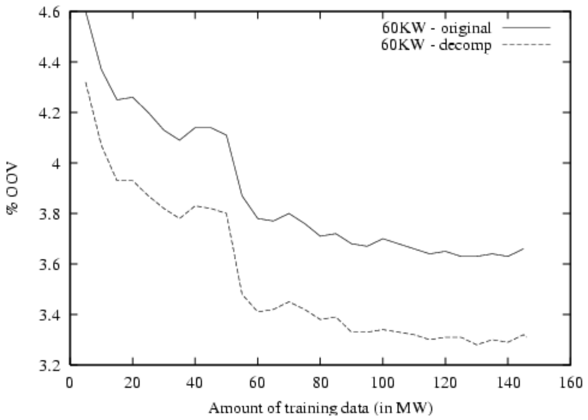

Figure 2: Effect on OOV rate

3.3. Proper Names and Acronyms

Proper names and acronyms deserve special attention in speech recognition development for Spoken Document Retrieval. They are important, information-carrying words but, especially in the broadcast news domain, also often out-of-vocabulary and therefore a major source of error. Because of this, it is important to predict as good as possible which proper names and acronyms are most likely to occur in the test data.

3.3.1. Selection of proper names and acronyms

In general proper names and acronyms are selected like any other word according to their frequencies in the development data. As a result, almost 28% of our 65K lexicon consists of proper names and acronyms. We did a few experiments to see how well frequency statistics can model the occurrence of proper names and acronyms. Given a 65K lexicon based on the PDB data set, we removed different amounts of proper names and acronyms according to a decision criterion and replaced them by words from the overall word frequency list, thus creating new 65K lexicons. To measure lexical coverage and OOV rates, we took the training data itself and, since we do not have accurate transcriptions of broadcast news, a test set of 35000 words of teletext subtitling information from January 2001 broadcast news, as a rough estimate of the actual transcriptions of broadcast news shows. In Table 5, lexical coverage and OOV rates of these lexicons are listed.

Corpus	amount	cov	OOV
PDB	27,9%	97,5%	2,5%
PDB	0%	91,4%	8,7%
PDB	15%	97,2%	2,8%
PDB	3,8% ($N > 100$)	96%	4,1%
PDB	0,7% ($N > 500$)	94,2%	5,9%
PDB	0,3% ($N > 1000$)	93,3%	6,7%
BN	27,9%	96,4%	3,6%
BN	0%	91,7%	8,3%
BN	15%	96,3%	3,7%
BN	3,8% ($N > 100$)	95,5%	4,5%
BN	0,7% ($N > 500$)	94,1%	5,9%
BN	0,3% ($N > 1000$)	93,6%	6,4%
BN	PDB 63K lower+BN 2K upper	94,4%	5,6%
BN	PDB + P&A	96,5%	3,5%

Table 5: Lexical coverage and OOV rates of 65K lexicons created with different amounts of proper names from different sources. PDB stands for the Persdatabank2000 subset, BN for the teletext broadcast news transcripts, N means frequency of occurrence in the training data and P&A Proper Names and Acronyms of December 2000.

Table 5 shows that selecting proper names and acronyms like any other word according to their frequencies in the development data, works very well with a lexical coverage of 97,5% and 96,4% for the PDB and BN data respectively. The next step was removing a part of the proper names and acronyms and replacing them by normal words from the word frequency list. Nor selecting only 15% instead of 27,9% proper names and acronyms, nor selecting only those with a frequency of occurrence of at least N (where N was 100, 500 or 1000), nor removing all proper names and acronyms, improved performance in lexical coverage on both test sets.

3.3.2. *Time dependency of proper names and acronyms*

In the broadcast news domain, it may be expected that recency is more important for proper names and acronyms than for normal words; i.e. names from people and firms may come and go relative to news events. Therefore one should expect that names and acronyms that recently popped up in broadcast news, will still be important in a short period thereafter. To investigate this, we took the 2000 most frequent proper names and acronyms from teletext subtitling information of the last quarter of 2000 and the remaining words from the Persdatabank2000 subset, and tested this lexicon on the 2001 broadcast news test set. The decrease in lexical coverage from 96,4% to 94,4% shows that the assumption above did not apply for our data. In a second experiment only proper names and acronyms from the preceding month were added that have a word frequency greater then 2. With this method lexical coverage improves somewhat, but not significantly (z=0.839). More experiments have to be done to draw stronger conclusions about recency effects on the selection of proper names and acronyms. For the time being, proper names and acronyms will be treated as normal words, selecting them according to their frequency only.

4. Conclusion and Summary

We have described the collection and preparation of a training collection for language modelling of the Dutch speech recognition system that will we deployed in Spoken Document Retrieval tasks. Although a collection of 152M words is only a start ("there's no data like more data"), it provides a fair basis for the envisaged research. We have been able to reduce the initially large amount of distinct words in our collection dramatically by applying several normalization procedures and by decompounding frequent compound words. Furthermore, our decompounding experiments showed interesting results when full and partial decompounding procedures were applied. The reported results seem to indicate that a partial decompounding procedure could improve lexical coverage when infrequent compounds are to be expected in the training data, whereas a full decompounding procedure should yield better results given training data that is lexically less refined. An optimal selection of proper names and acronyms, is something we need to look into a little further. In our experiments, best lexical coverage was obtained by ignoring their special nature, just selecting them like any other word according to its frequency in the training data. Attempts to improve the lexicon by calling in recent text data to predict the occurrence of proper names and acronyms, was hardly successful. This could well be due to the fact that the time span we labeled 'recent' was not recent enough.

5. References

[1] DRUID Demonstrators. `http://dis.tpd.tno.nl/druid/public/demos.html`.

[2] DRUID Project Homepage. `http://dis.tpd.tno.nl/druid/`.

[3] ECHO Project Homepage. `http://pc-erato2.iei.pi.cnr.it/echo/`.

[4] G. Adda, M. Jardino, and J. Gauvain. Language Modelling for Broadcast News Transcription. In *Eurospeech'99*, pages 1759–1762, Budapest, 1999.

[5] M. Adda-Decker, G. Adda, and L. Lamel. Investigating text normalization and pronunciation variants for German broadcast transcription. In *ICSLP'2000*, pages 266–269, Beijing, 2000.

[6] M. Adda-Decker, G. Adda, L. Lamel, and J.L. Gauvain. Developments in Large Vocabulary, Continuous Speech Recognition of German. In *IEEE-ICASSP*, pages 266–269, Atlanta, 1996.

[7] M. Adda-Decker and Lamel L. The Use of Lexica in Automatic Speech Recognition. In F. van Eynde and D. Gibbon, editors, *Lexicon Development for Speech and Language Processing*. Kluwer Academic, 2000.

[8] J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference*, pages 107–129, Washington, 2000.

[9] Ronald RosenFeld. Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data. In *Eurospeech-95*, pages 1763–1766, 1995.