Evaluation of Lexical Models for Hungarian Broadcast Speech Transcription and Spoken Term Detection

Balázs Tarján*, Péter Mihajlik*,**, András Balog** and Tibor Fegyó*,***
*Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
**THINKTech Research Center, Hungary
***AITIA International Inc., Hungary
tarjanb@tmit.bme.hu, mihajlik@tmit.bme.hu, balog@thinktech.hu, tfegyo@aitia.ai

***Abstract*—In this paper, we re-evaluate morph (data-driven subword) and word lexical models used for large vocabulary continuous speech recognition of agglutinative languages. Since such speech recognition systems are applied mostly for information retrieval purposes we use evaluation metrics accordingly. Standard 3-gram language model with one million words vocabulary is used for words whereas statistical morph-based models are applied with smaller vocabularies and with higher order of n-gram models. Fostering real life applicability, the computational time and memory usage of the various approaches is kept below real-time and 1.5 GB, respectively. The lexical modeling approaches are tested on Hungarian Broadcast News and Broadcast Conversation speech. In our setup, although word-based models outperformed morph-based ones in terms of both word error rate and spoken term detection measures, a search-cascade of the word and morph approaches improved the latter results significantly.**

Keywords – speech recognition, LVCSR, agglutinative languages, broadcast news, broadcast conversation, spoken term detection

I. Introduction

Large vocabulary continuous speech recognition (LVCSR) of morphologically rich languages – like Hungarian – can be challenging due to rapid vocabulary growth. Huge lexicons, high out of vocabulary (OOV) rate and sparse data for language modeling are the well-known symptoms of word-based lexical approaches. Many efforts have been made to overcome these difficulties – the most successful techniques apply various vocabulary decomposition methods [1]. Stem-ending, grammatical morphs and statistical morphs performed significantly better than the word baseline for several languages in speech recognition and speech retrieval tasks [1,2,3]. However, some questions have still remained unanswered. Next, we briefly overview the task investigated, the problems and the related work.

Our task to be solved is real-time, resource efficient spoken term detection (STD) in Hungarian Broadcast speech. The aim is to implement a system that is able to monitor ad-hoc, unrestricted index terms “on the fly” in Hungarian broadcast speech on multiple channels. Since no prior publication dealing with the same topic is known for the authors, other language and other task results are considered as related work. At first, it is clear that the application of information retrieval (IR) methods on the output of an LVCSR system is the most promising approach [4]. One of the most important questions is what kind of word- or subword-based lexical modeling technique performs best in LVCSR if the evaluation metrics are STD measures. Earlier works [3,5] show the advantage of morph-based approach in Hungarian LVCSR in terms of word error rate (WER), but our recent investigations suggest that this advantage can significantly decrease with larger training databases and with smaller vocabulary growths [6]. On the other hand, WER may not be a good indicator of IR performance since i), named entities that can be critical in IR are severely underrepresented in WER and ii), word stems are more appropriate than whole words in retrieval tasks for the Hungarian language [7]. So, the evaluation of STD performance of lexical modeling approaches on Hungarian Broadcast Speech is a missing, real-life oriented work to be completed.

Considering speech information retrieval results of other agglutinative languages, recent advances made for Turkish language should be emphasized. A reason for selecting Turkish is that it is one of the most similar languages to Hungarian in terms of typical vocabulary growth and morphological structure [3]. Spoken term detection is introduced for Turkish Broadcast News in [8]. The statistical morphs outperformed the word baseline both in terms of WER and IR measures. Sub-word based language modeling improved WER from 23.2% to 20.4% and ATWV from 64.5% to 75.8%. The best STD results are obtained using word-morph-cascade approaches. In the vocabulary-cascade technique (79.5% ATWV), word-based system is used for intra vocabulary (IV) queries and morph-based one for the OOV search terms. Similar results were achieved with a search-cascade (81.1% ATWV), where index terms not found with the word-based system are searched with the morph-based approach [8]. The evaluation followed the NIST 2006 recommendations that apply whole word matching. This technique, however, is impractical for agglutinative languages since it can cause high number of false negative STD results. It is clear that if the English index term "*European Union*" is considered, then the strings in the recognizer's output "*in the European Union*" or "*for the European Union*" mean true positive results. In Hungarian, the translated index term would be "*Európai Unió*" and the recognized strings "*az Európai Unióban*" or "*az Európai Uniónak*". In this case both relevant and correct result would be missed. Stemming and/or other auxiliary techniques should therefore be applied for agglutinative languages in STD evaluations. The application of stemming though can re-order various lexical approaches in terms of STD performance – that is why we investigated this phenomenon, as well.

In sum, the novelties of our study are that i), STD evaluation is based on stems ii), novel subword-based lexical modeling approach is introduced in order to improve detection of named entities iii), no vocabulary cutoff is applied to enable fair comparison of lexical modeling approaches. For the authors no prior publication dealing with morphologically rich languages is known that applied any of the above.

II. Concept and Tasks

Our primary aim is to evaluate LVCSR lexical model in terms of spoken term detection under real-life conditions. Practically speaking, the task is to design a cost-efficient, low-latency, real-time STD system for monitoring Hungarian broadcast speech. Both Broadcast News and Broadcast Conversations speech should be followed. By default, the speech genre is not known a priori. To achieve fast operation, a one-pass decoding and spoken term detection strategy is applied.

Text pattern matching and morphological analysis based IR methods are run on the 1-best output of the speech recognizer, thus no lattice or word confusion network generation and search is performed. This limits STD evaluations as only binary decision can be made for the index terms. However, in this way decoding speed can be very high even if huge vocabularies are used (see Fig. 1). The evaluated real-time STD system is an enhancement of the Mindroom[1] public Hungarian broadcast speech retrieval system.

A. Recognition Tasks

The lexical models are evaluated on *Broadcast News* (BN) and *Broadcast Conversations* (BC) type of speech data collected from the archives of the Mindroom system. This broadcast speech retrieval system monitors five Hungarian TV channels and one radio station. For parameter optimization purposes development (dev) sets are defined, and for evaluations, independent evaluation (eval) sets are collected sampling the speech data representatively. All test sets are constructed from around 2 minutes long samples extracted from monitored TV programs. (see Table I.)

B. Training Data

Manual transcriptions were available only for 50 hours of broadcast – mostly conversational – speech containing 370k words altogether. Two additional training text corpora have been collected from the website of the TV channels, one from news columns (27M words) and one from tabloid columns (14M words) of the corresponding websites. There is no time overlap between any training and test data.

TABLE I. Basic characteristics of the test sets

Test sets	***BC dev***	***BC eval***	***BN dev***	***BN eval***
Length [min]	36	52	31	49
# of words	5108	7524	3988	6116
OOV [%]	1.76	1.58	1.76	1.75
Stem OOV [%]	0.84	0.7	0.72	0.9
Word PPL	848	1055	928	885

III. Evaluation System Description

In the following, the LVCSR-based STD system used to evaluate the lexical approaches is introduced.

A. Lexical modeling approaches

Speech recognition and the related IR performance of our system are evaluated by using three different lexical modeling techniques – a classical word-based and two data-driven morph-based ones. Note that experiments with grammatical morphs and with stem+ending lexical models were also conducted, but as their speech recognition performance were not comparable to the best performing lexical approaches they are not presented here.

1) Word-based approach

According to the standard word-based approach, word tokens are served as basic elements of the language model. Aiming at the maximization of keyword detection measures, we intended to return named entities in their original, capitalized form. Thus, we did not lower-case the training text in general but the sentence beginning characters. In order to avoid unwanted case conversions, a grammatical parser [7] was used for named entity recognition.

2) Case insensitive morph-based approach (MB-CI)

This morph-based lexical modeling approach is based on a widely used unsupervised vocabulary decomposition technique, called Morfessor Baseline (MB) [9]. As usual, every word in the training text (except acronyms) is mapped into its lower case equivalent. The word-to-morph dictionary and the word boundary reconstruction technique is the same as in [3] and [5]. The core word list given to the unsupervised MB algorithm contains all word forms – only acronyms are filtered out. This technique performed best on other Hungarian LVCSR tasks in terms of speech recognition accuracy [3,5].

3) Case sensitive morph-based approach (MB-CS)

An evident drawback of the previous morph-based approach is that it is not able to return named entities in their original, capitalized form. In order to overcome this problem, a modified morph-based lexical modeling technique was worked out. In contrast with the case insensitive method, here the letter cases in the word lexicon are retained, and not only acronyms, but all capitalized words are removed from the core word list. The collection of these exception words is then stemmed and the resulting suffixes are added to the core words, as well. Thus, after statistical decomposition, only lower case words

[1] http://mindroom.hu

and the suffixes of capitalized words are mapped into morph sequences, whereas capitalized word stems and acronyms remain intact.

B. *Language models*

For each training corpus (370k words transcriptions, 27M words news, 14M words tabloid corpus) a separate language model is built by applying modified interpolated Kneser-Ney smoothing implemented in the SRILM toolkit [10]. This toolkit is also used to perform linear interpolation of the previous three language models. Interpolation weights are determined by optimizing word perplexities on the merged development sets of BN and BC tasks – thus, the same "global" language model is used for both tasks. 3-gram model is built for the words and 4-gram models for the morphs (singleton 3- and 4-grams are discarded in both cases). Entropy-based pruning [11] was applied only for the word-based model in order to reduce its size – and this way its memory consumption – to the level of the morph-based models.

C. *Pronunciation model*

Simple grapheme-to-phoneme rules and exceptions are applied on each lexicon separately in order to obtain word-to-phoneme and morph-to-phoneme mappings. Automatic phonetic transcription of both morphs and words can result in pronunciation errors at morph boundaries; however, according to our former experiences this kind of error occurs rarely [5]. A manually collected dictionary containing 5000 weighted alternative pronunciations is also used in order to support the detection of words pronounced differently from the standard Hungarian way.

D. *Acoustic model*

Speaker independent, cross-word triphone models are trained using decision trees and MMI (Maximum Mutual Information) estimation with the HTK toolkit [12]. For both the BC and BN tasks the same acoustic model is trained on 50 hours of broadcast data. Number of states is about 5000 and 13 Gaussians are used per state. The feature type is MFCC (Mel Frequency Cepstral Coefficients) + log Energy with delta and delta-delta, calculated on 8 kHz bandwidth speech, and blind channel equalization [13] is also applied.

E. *Decoding*

To avoid biasing the comparison of lexical models no vocabulary cutoffs are made at all. This results in very large word vocabulary size (see Table II) which challenges conventional decoders, especially if real-time operation is prescribed. The WFST (Weighted Finite State Transducer) framework offers an attractive solution to make the search space as compact as possible [14], thus, by default, we applied our WFST decoder called as VOXerver-flat[2] for this task. VOXever-flat is an upgraded version of VOXerver providing faster operation and low memory consumption. The WFST recognition network construction and optimization scheme is the same as in [3]. To check the real-time applicability of our approach, Word-Accuracy – RTF (Real Time Factor) curves of word-based systems are illustrated in Fig. 1., where performance of VOXerver-flat can be compared to another recently developed WFST-based decoder called Juicer Lite[3] and to a conventional LVCSR decoder, HDecode[4].

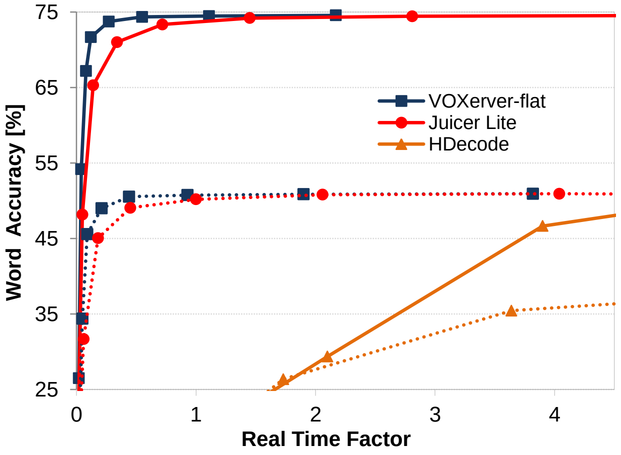

Figure 1. Performances of VOXerver-flat and Juicer Lite WFST decoding approaches with word-based lexical models compared to the HDecode-word baseline. Results on Broadcast News (BN) data are denoted with solid curves, whereas Broadcast Conversation (BC) results are illustrated with dotted curves.

As can be seen in this figure, both WFST decoders significantly outperformed the conventional LVCSR decoder (HDecode). The difference between the WFST decoders (Juicer Lite vs. VOXerver-flat) is smaller but VOXerver-flat is still about 2-3 times faster than the Juicer Lite in the under-saturated regions. The speed enhancement of VOXerver-flat can be due to the very compact HMM-WFST representation (almost half memory demand than of the Juicer Lite) which enables a more optimal CPU cache usage.

In the further experiments RTF's of morph- and word-based systems are adjusted to be nearly equal using standard pruning techniques. RTF values are set close to 0.7 – measured on the same 2.26 GHz Intel Xeon processor. Memory consumption of decoding is also balanced for the different lexical approaches in the range of 1.1-1.4 GB when using the VOXerver-flat engine. (HDecode required about 600MB of memory, whereas consumption of Juicer Lite was in the range of 1.7-2.1 GB)

F. *Spoken term detection*

As it is mentioned earlier, a simple spoken term detection approach is used in the experiments. To enable real-time, low-latency spoken term detection, the *textual* recognition output is retrieved from decoder in every 150ms. If a word is constant for at least 3 cycles the corresponding time segment can be indexed instantly. In the experiments both the reference index terms and recognized words are stemmed and the corresponding string matching is observed. In the real system stemming, however, could be too slow, so lexical and altered index stems may be matched with word beginning characters. This approximation can introduce some errors but ensures real-time operation. In this paper we strictly evaluate IR performance with application of a grammatical stemmer [7] both on the indices and on the LVCSR transcriptions.

[2] http://www.aitia.hu/voxerver

[3] http://juicer.amiproject.org/juicer/

[4] http://htk.eng.cam.ac.uk/extensions/

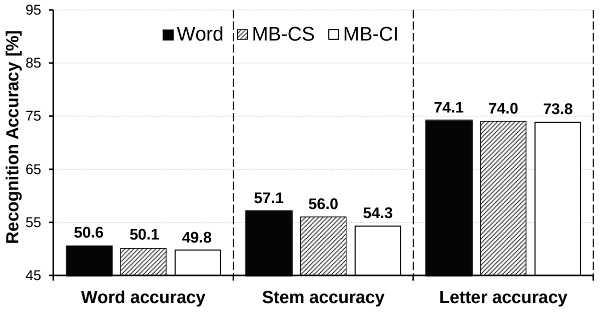

Figure 2a. LVCSR results on Broadcast Conversation (BC) evaluation set with various lexical models

Although their real-time implementation can be questionable, both the vocabulary- and search-cascade techniques ("VCasc" and "SCasc" in Table III and IV) described in [8] are evaluated cascading the word- and MB-CS-based systems. In vocabulary-cascading IV queries are retrieved from the word-based index, whereas OOV queries are searched in the morph-based index. This approach relies on the assumption that word-based recognizer generally performs better on IV words, whereas morph-based index is better on OOV terms. On the other hand in search-cascading every query is first searched in the word-based index. Morph-based index is searched only in the case if there is no match in the word-based index for the term.

IV. Experimental Results

A. Speech recognition results

In order to enable fair comparison of different lexical modeling approaches, all LVCSR results are evaluated in case insensitive way. Word Error Rates (WER) of lexical modeling approaches on the development and evaluation sets are displayed in Table II. For further experiments only the evaluation sets are used and accuracies are given.

Since in case of morphologically rich languages word accuracy can show a pessimistic picture of the speech recognition performance [1], letter accuracy is also calculated for the evaluation setups. Whitespace tokens between words are modeled by a dedicated letter in our letter accuracy calculation. Considering that LVCSR system is used for solving STD task, we measure stem accuracies, as well. LVCSR results on the BC and BN evaluation sets can be seen in Fig. 2a and 2b, respectively. In short, the word-based system performs best nearly in all measurements.

TABLE II. Vocabulary sizes and WER results [%] of the applied lexical models on all test sets

Lex. Model	*Vocab size*	*BC dev*	*BC eval*	*BN dev*	*BN eval*
Word	1M	45.4	49.4	25.9	26.3
MB-CS	230k	46.2	49.9	26.0	26.8
MB-CI	79k	46.3	50.2	26.2	26.5

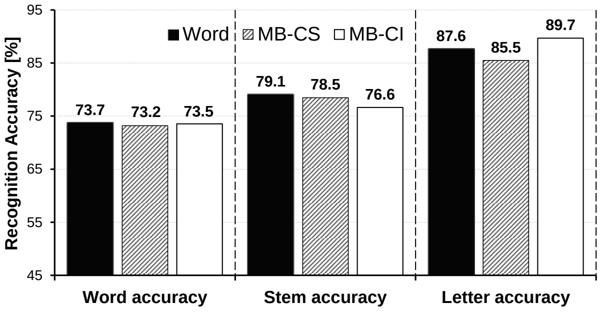

Figure 2b. LVCSR results on Broadcast News (BN) evaluation set with various lexical models

B. Spoken term detection results

TABLE III. STD results of various single-pass and cascade approaches for NE and random word queries on BC and BN tasks using classic metrics [%]

Test / Lex. Model	NE q.s.	Random query set		
	F1	*Precision*	*Recall*	*F1*
BC Word	50.1	66.8	61.3	61.7
BC MB-CS	48.1	64.0	58.8	59.2
BC MB-CI	17.9	59.4	54.0	54.6
BC VCasc	49.9	66.6	60.8	61.3
BC SCasc	54.3	69.2	63.7	64.1
BN Word	60.1	77.8	75.8	75.6
BN MB-CS	59.4	77.4	74.4	74.7
BN MB-CI	20.5	73.1	70.5	70.7
BN VCasc	60.1	78.2	76.1	76.0
BN SCasc	64.2	81.2	78.9	78.8

TABLE IV. ATWV results of various single-pass and cascade approaches for NE and random word query sets (detailed over IV, OOV, OOV stem queries, as well) on BC and BN tasks [%]

Test / Lex. Model	NE q.s.	Random query set			
		All	*IV*	*OOV*	*OOV stems*
BC Word	43.7	47.3	48.8	20.1	-
BC MB-CS	42.3	44.7	46.4	15.0	0
BC MB-CI	14.7	39.8	41.4	12.0	0
BC VCasc	43.5	46.9	48.8	15.0	0
BC SCasc	47.7	49.3	50.7	24.2	0
BN Word	57.0	68.1	70.6	16.6	-
BN MB-CS	55.6	67.4	69.4	27.6	4.8
BN MB-CI	19.0	63.8	65.6	27.6	4.8
BN VCasc	57.0	68.5	70.6	27.6	4.8
BN SCasc	60.7	71.1	73.2	30.3	4.8

The performance of our STD system is characterized with classical IR metrics (precision, recall, F-measure), and with ATWV (Actual Term Weighted Value, β=1000), which is a metric developed especially for the evaluation of STD tasks [15]. All values are averaged over the corresponding query set. For each task (BC, BN) *two* single-term query sets are defined. The first "NE" query list contains all the spoken Named Entities (NE) of the evaluation sets, 150 queries for BC, 200 queries for BN task. The second "Random" query list consists of words randomly chosen from the manual transcript of the corresponding evaluation set, 840 queries for BC, 970 queries for BN task.

V. Discussion

Speech recognition results (Table II., Fig. 2.) suggest that word-based recognizer can outperform morph-based one if sufficient amount of textual training data is available. Looking further, the results confirm our theory about speech recognition improvements due to vocabulary decomposition. That is, according to the previous results summarized in [3] and [6], vocabulary decomposition can be an effective mean to reduce data sparseness for language modeling. A weakness though is that an overhead is required for the restoration of word forms. As the amount of training data and the vocabulary increase, the standard word-based technique may outperform sub-word based approaches – as in our case. It is still an open question whether this finding is applicable for languages with much faster vocabulary growths such as Finnish or Estonian.

As can be seen in Table III and IV, among the single-pass STD approaches, the word-based system performs best in each setup. As for the BC task, surprisingly, the classical word lexical model was found superior not only for intra vocabulary queries but also for OOV terms due to stemming. As a consequence using vocabulary-cascade resulted in no significant improvement either for BC or BN task.

Although morphs performed poorer than words in terms of WER and STD measures, sub-word based recognition can supplement word-based one due its ability to handle data sparseness. Accordingly word-morph *search-cascade* improved the STD results significantly ($p < 0.0001$) over the single pass word-based system. The fact that morph-based technique in a search-cascade can increase not only OOV but IV STD scores suggests that there are terms which are not sufficiently modeled in word-based language models and only can be detected with morph-based models.

VI. Conclusions

We have introduced techniques for fast and resource efficient STD in Hungarian broadcast speech. The morphological richness of Hungarian was addressed by data-driven vocabulary decomposition methods as well as by extra-large word vocabulary. The previously best performing [3,5,6] morph-based lexical approach (MB-CI) reached the lowest letter error rate on the Broadcast News task; however, its STD performance was the poorest in all tests. We observed that STD measures correlate mostly with stem accuracies of the LVCSR results. The baseline morph-based approach was extended in order to handle named entities correctly (MB-CS) that improved significantly the STD performance.

In our experiments none of the morph-based approaches could outperform the word-based technique. Even OOV terms were detected better with the word-based system due to stemming. It is acknowledged that stemming is not common in standard STD evaluations but we argue that its application in case of agglutinating languages bring the evaluation closer to the English language case. Considering word-morph cascading techniques only *search-cascade* [8] was able to improve STD scores due to the utilization of morph-based index for terms not sufficiently modeled in word-based system.

As for future work, we plan to make the STD evaluation using multiple hypothesis speech recognition output with advanced thresholding techniques [8]. Furthermore, growing and pruning techniques [16] for higher order of (morph) n-gram models are to be investigated in order to fully exhaust the potential of subword-based approaches.

Acknowledgment

This research was supported by the KMOP 1.1.3-08/A-2009-0006, by the KMOP-1.1.1-07/1-2008-0034, by the TÁMOP-4.2.2-08/1/KMR-2008-0007 and by the TÁMOP-4.2.1/B-09/1/KMR-2010-0002 projects. The authors also wish to thank Dr. Géza Németh for his kind support towards this work.

References

[1] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimaki, J. Pylkkonen, T. Alumae, and M. Saraclar, "Unlimited vocabulary speech recognition for agglutinative languages," in HLT-NAACL, New York, USA, June 5-7, 2006.

[2] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar, "Turkish Broadcast News Transcription and Retrieval," IEEE Trans. Audio Speech Lang. Proc., 17(5):874-883, 2009.

[3] B. Tarján and P. Mihajlik, "On Morph-based LVCSR Improvements," in SLTU 2010, Penang, Malaysia, pp. 10-16.

[4] D. R. H. Miller et al., "Rapid and accurate spoken term detection," in INTERSPEECH-2007, pp. 314-317.

[5] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh, and T. Fegyó, "Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task," IEEE Trans Audio Speech & Lang. Proc., vol.18, no.6, pp.1588-1600, Aug. 2010

[6] L. Tóth, B. Tarján, G. Sárosi, and P. Mihajlik, "Speech Recognition Experiments with Audiobooks," Acta Cybernetica-Szeged, 19(4): pp. 695-713. (2010)

[7] V. Trón, L. Németh, P. Halácsy, A. Kornai, Gy. Gyepesi, and D. Varga, "Hunmorph: open source word analysis," in Proc. ACL 2005 Software Workshop, pp. 77-85.

[8] S. Parlak and M. Saraclar, "Spoken term detection for Turkish Broadcast News," in ICASSP 2008., pp.5244-5247

[9] M. Creutz and K. Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.," in Comp. and Inf. Sci., report A81, HUT, March 2005.

[10] A. Stolcke, "SRILM – an extensible language modeling toolkit," in Proc. Intl. Conf. on Spoken Language Processing, pp. 901–904, Denver, 2002.

[11] A. Stolcke, "Entropy-based pruning of backoff language models," in Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, USA, 270–274

[12] S. Young et al., The HTK book. (for HTK version 3.4.), 2006

[13] L. Mauuary, "Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition," in Proc. EUSPICO'98, Vol.1, pp. 359-363, 1998

[14] M. Mohri, F. Pereira, and M. Riley, "Weighted Finite-State Transducers in Speech Recognition," Computer Speech and Language, 16(1):69-88, 2002.

[15] NIST, "The Spoken Term Detection (STD) 2006 Evaluation Plan," http://www.nist.gov/speech/tests/std/, 2006.

[16] V. Siivola, T. Hirsimäki, and S. Virpioja, "On Growing and Pruning Kneser-Ney Smoothed N-Gram Models," IEEE TASLP, 2007, pp.1617-1624.