Studies on Training Text Selection for Conversational Finnish Language Modeling

Seppo Enarvi and Mikko Kurimo

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
seppo.enarvi@aalto.fi

Abstract

Current ASR and MT systems do not operate on conversational Finnish, because training data for colloquial Finnish has not been available. Although speech recognition performance on literary Finnish is already quite good, those systems have very poor baseline performance in conversational speech. Text data for relevant vocabulary and language models can be collected from the Internet, but web data is very noisy and most of it is not helpful for learning good models. Finnish language is highly agglutinative, and written phonetically. Even phonetic reductions and sandhi are often written down in informal discussions. This increases vocabulary size dramatically and causes word-based selection methods to fail. Our selection method explicitly optimizes the perplexity of a subword language model on the development data, and requires only very limited amount of speech transcripts as development data. The language models have been evaluated for speech recognition using a new data set consisting of generic colloquial Finnish.

1. Introduction

Finnish language has a colloquial variant that differs from the formal literary Finnish substantially. While clearly pronounced literary Finnish can already be recognized with high precision, current ASR systems are unable to recognize conversational Finnish, because there has not been any training or evaluation data available.

With regard to a speech recognizer, the set of phonemes is the same in both language varieties, but the difference in vocabulary and grammar is clear [1], so we have started the research on colloquial Finnish NLP by collecting text data. The relevance of the collected text for speech recognition has been evaluated with Aalto speech recognizer. In addition to speech recognition, the data is valuable for other tasks such as machine translation as well, because Finnish language communication more and more includes colloquial characteristics [2].

So far there are no statistical language models that would cover colloquial Finnish. Finnish conversations in e.g. Internet are written down phonetically, often including phoneme reductions and compounding, suggesting that on-line discussions would offer useful data for language modeling.[1] While there are huge amounts of data available, it is important to select only what is useful for the modeling task. The irrelevant n-grams increase confusability and computational burden in language models. Irrelevant data also makes analysis such as discovery of morphemes and word classes error-prone and computationally more intense. For these reasons we have evaluated speech recognition errors and language model perplexities, as well as the reduction in data size.

Related research has been carried out earlier in the context of adapting an out-of-domain language model with in-domain data. A popular approach has been to train an in-domain language model and select text segments with low perplexity [3]. Klakow trained language models from out-of-domain data, computing the change in in-domain perplexity, when a text segment is removed from the training data [4].

Sethy et al. used relative entropy to match the distribution of the filtered data with the in-domain distribution [5]. Instead of scoring and filtering each text segment individually, they select text segments sequentally, adding a new segment to the selection if it reduces relative entropy with respect to the in-domain data. The algorithm was later revised to use a smoothed version of the Kullback-Leibler distance that uses a tunable smoothing parameter [6], with improved results.

Moore and Lewis used formal reasoning to show that if the selection method is based on the probability (in terms of cross-entropy or perplexity) given by an in-domain language model to the training text segment, one should compare the probability to the probability given by an out-of-domain language model [7]. They computed the cross-entropy of each text segment according to an in-domain language model and an out-of-domain language model, and used the difference between the two cross-entropies as the selection criterion.

From the above approaches the one proposed by Klakow requires the least amount of in-domain development data, since models are estimated only from the out-of-domain data. At the time we had very little in-domain development data of conversational Finnish (we used a set of 1047 utterances in these experiments), so this was the only applicable approach. The method may become computationally demanding since it

[1] The most notable difference between transcribed speech and written conversation is that disfluencies are usually omitted in writing.

requires training as many language models as there are text segments, but the computation can be done in parallel.

Another line of research has used information retrieval techniques to select in-domain documents. Term frequency–inverse document frequency (tf-idf) is a popular measure of document similarity. After constructing a vector representation of each document, it is efficient to find documents that are similar to a query string. Mahajan et al. proposed to use it for language model adaptation based on current recognition history [8].

We have collected Internet conversations using Google search, and by crawling Finnish discussion sites. The obtained text segments are scored, and the worst scoring segments are pruned. The threshold score for pruning text segments is found automatically, so as to minimize the perplexity of the resulting language model on a held-out data set. The unlimited vocabulary presents challenges in using perplexity for scoring and for finding the pruning threshold. The perplexity optimization is possible only with a subword language model. We have also collected a new set of transcribed Finnish conversations for development and evaluation purposes.

The next section discusses the challenges posed by the unlimited nature of Finnish vocabulary. Section 3 presents our new development and evaluation data. In Section 4 we explain how we have collected web data for language modeling. Section 5 describes how we have performed our evaluations, and the results are given in Section 6. Finally, Section 7 draws conclusions.

2. Vocabulary in conversational Finnish speech recognition and perplexity computation

The highly agglutinative nature of Finnish language makes it difficult to create an exhaustive vocabulary for speech recognition. Creutz et al. show a comparison of vocabulary growth across different languages [9]. While conversational speech is generally thought to be less diverse than planned speech, there are no less word forms used in conversational Finnish text than in a similar amount of literary Finnish. The reason is that the phonetic variation in conversational Finnish is translated into new vocabulary.

Finnish orthography is very close to phonemic, meaning that written letters generally correspond to spoken phonemes. In informal conversations, phonetic variation is also often reflected in writing, even to the extent that sandhi is expressed in written form. For example, “en minä tiedä” is literary Finnish, and can be translated “i don’t know”. Reduced forms of the same expression are used in spoken conversation, but often in textual communication as well:

en mä tiedä
en mä tiiä
emmä tiiä

The situation is different from English, where there is generally only one way to spell each word, even though several different pronunciations exist. When having a spoken conversation, one could actually utter /aɪ doʊnt noʊ/, /ədənoʊ/, or /dʌnəʊ/, but in any of these cases one would probably write “i don’t know”, if the conversation was textual.

A comparison of vocabulary growth in Finnish and English Internet conversations and formal texts is shown in Figure 1. The formal English plot has been created from newspaper corpora. The formal Finnish data is literary Finnish from books and newspapers. The conversational Finnish text is Internet conversations from Suomi24 discussion site, covering many different topics. The conversational English is gathered from the web by searching text related to topics in meeting transcripts from CMU, ICSI, and NIST [10]. Web texts were processed using normalization scripts. It should be noted that the quality of text normalization may vary, as well as the degree to which the web data sets are spontaneous and colloquial.

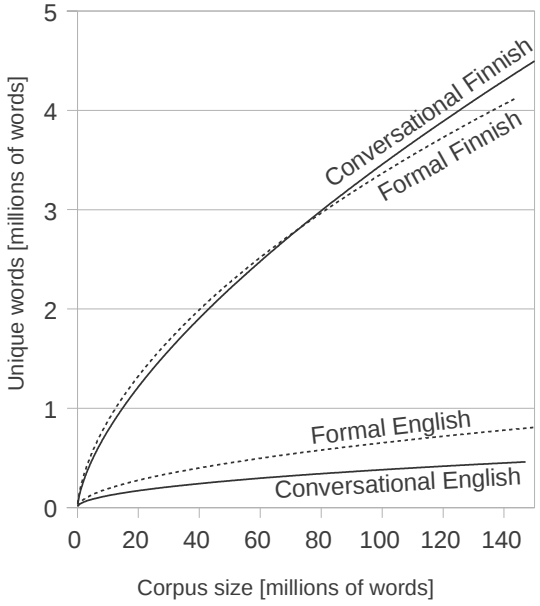

Figure 1: Vocabulary growth, when all the encountered words are added to the vocabulary, on newspaper-style formal text and Internet conversations

The curves show that, as expected, vocabulary growth in formal English is clearly faster than that in English conversations. However, in Finnish Internet conversations vocabulary grows at a similar pace to, and eventually exceeds that of formal Finnish.

Another comparison was made to see how the vocabulary growth affects OOV rates, by using an independent test set from each category. Figure 2 illustrates the percentage of words in the corresponding test set that are missing from the training set, for growing amounts of training data. The training data is the data used to plot Figure 1. The formal English test data was transcribed broadcast news speech, and the formal Finnish was planned literary Finnish from the SPEECON [11] corpus. The conversational test data sets

were transcribed conversations, omitting hesitations. The high OOV rates on transcribed Finnish conversations suggest that the vocabulary growth in Finnish Internet conversations is not just a result of poor text normalization or clean up.

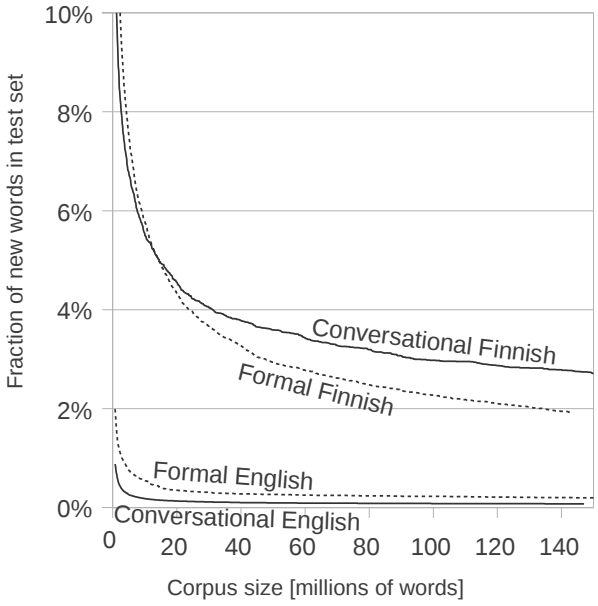

Figure 2: Development of OOV rate, when all the encountered words are added to the vocabulary, on newspaper-style formal text and Internet conversations

The standard approach for unlimited vocabulary Finnish language speech recognition has been to use statistical morphs as the basic language modeling unit, instead of words [12]. It seems that statistical morphs obtained by direct application of Morfessor Baseline [13] to the word list do not model conversational Finnish well. The reason may be insufficient quality or quantity of training data, or the pronunciation variation behind new word forms. Factored language models [14] is one way to alleviate the vocabulary size issue, but at the moment there are no tools for extracting meaningful factors from colloquial Finnish word forms. Development of such tools would be extremely difficult because of the numerous ways in which phonetic variation can alter the words.

We tried conversational speech recognition with morph-based models, but so far there was no improvement over word models in terms of word error rate. However, the perplexity computations in the text selection algorithm have been performed using morph models. The reason is that there are so many OOV words that we need reliable estimates also for n-grams containing OOV words. Even though the initial morph models are not yet sufficiently good for ASR, they seem to offer a reasonable approximation for perplexity computation.

Language model perplexity is generally computed either including only those words that occurred in the training data, or using an open-vocabulary language model, i.e. one that contains the unknown word token <UNK>. The probability for unknown words is obtained by replacing the most infrequent words in the training data with <UNK>, or by discounting the observed unigram probabilities.

If one chooses to use a closed vocabulary, and compute perplexity only on in-vocabulary n-grams, the perplexity value will increase when the number of OOV words decreases. This makes perplexity optimization in Finnish difficult, since we do not know if we should prefer low perplexity or low OOV rate. This problem is easily overlooked with English language data, because the percentage of OOV words stays constant enough not to play a significant role in determining the perplexity value.

We did not find open-vocabulary language models to be a suitable solution either. The problem is that the selection algorithm is significantly affected by how the <UNK> probability is determined. The collected conversational Finnish text contains so many word forms that occur only once, that their probability mass alone gives a too high estimate for the OOV probability. Selection of text segments based on perplexity of such a model would prefer segments with high OOV rate.

3. Transcribed Finnish conversations

For development and evaluation data, we have transcribed Finnish conversations: five radio conversations from 13 different speakers, three podcast conversations from 5 different speakers, and recordings of 67 students discussing in pairs with headsets on. These conversations encompass a diverse set of speaking styles and topics, as the intention of this research is not to adapt statistical models to a specific topic or domain, but to collect generic colloquial Finnish data. The students were encouraged to discuss from any topic they could think of, although they were given 16 example topics. They could also use a web browser to find conversation topics from news sites. Only a portion of each conversation containing fluent conversation was selected for transcription. The discussions were entirely colloquial and very natural.

The conversations were divided into development and evaluation sets so that the same radio programs or speakers do not appear in both development and evaluation set. In total the evaluation set contains 44 minutes of audio, 541 utterances, and 17 different speakers. DEVEL1 development set contains 1047 utterances from 49 speakers, and DEVEL2 contains 445 utterances from 19 speakers.

Development data is required for filtering out irrelevant text. It may be used for both scoring text segments, and optimizing the rejection threshold, as explained in the next section. It is essential that in such case, different development data is used for text scoring and for finding the threshold score. Otherwise filtering will be too intense because of overfitting. When development data is not needed for text scoring, we have used the entire DEVEL1 and DEVEL2 data sets for optimizing the rejection threshold. Otherwise DEVEL1 has been used for scoring, and DEVEL2 has been held out for optimizing the filtering threshold. Both DEVEL1 and DEVEL2 sets were also included as training data for the acoustic model used in the speech recognition experiments.

4. Collecting and filtering web text for modeling Finnish language

4.1. Collected web corpora

Internet search engines are commonly used to query text for language modeling [10]. We collected several text corpora from the Internet, first using a script that extracts results from Google queries. The data set WEB1 was retrieved using devised 2-grams, 3-grams and 4-grams as query strings. The query n-grams were constructed from colloquial word forms, intentionally forming expressions that are used only in conversational Finnish.

A more systematic way to gather data set WEB2 was used. We extracted all the 3-grams from a transcribed radio conversation, and those that exist in a literary Finnish corpus were removed. The remaining 667 3-grams were used as search queries. We did not try other n-gram lengths, but 4-grams rarely return more than a few search results, and 2-grams are often too generic, returning even other than Finnish text. Surprisingly, WEB2 data did not improve recognition performance. Also, without a substantial amount of existing in-domain text, the amount of data obtained with this method was still small.

Data set WEB3 was extracted by copying the entire contents of a web site containing Internet Relay Chat (IRC) conversations. Data sets WEB4 and WEB5 were each collected by crawling a Finnish discussion site using Python libraries Scrapy and Selenium, and extracting every conversation. This turned out to be a fast method for obtaining large amounts of structured data.

4.2. Preprocessing web text

Extensive preprocessing was needed, before the web data could be used for language modeling. This included

- removal of non-textual items, such as hyperlinks, message board markup code, usernames, and smileys,
- expansion of abbreviations, numbers, punctuation marks, and such, and
- deletion of words that contain phoneme sequences that do not pertain to Finnish phonological rules.

Numbers do not carry information about pronunciation. We have simply expanded them as they are pronounced in literary Finnish. The sizes of the data sets after preprocessing are shown in Table 1.

Data set	Number of words
WEB1	767,669
WEB2	1,067,993
WEB3	562,426
WEB4	25,131,015
WEB5	46,258,268
DEVEL1	17,209
DEVEL2	8,755

Table 1: Data sets and their sizes after preprocessing

4.3. Text segment scoring

Data filtering starts by giving a numeric score to each text segment. Then segments whose score is below a threshold will be rejected from the training data. A shortcoming of this one-pass scheme is that every example of a common, short sentence receives the same high score, which may skew the distribution of the selected data too much towards frequent utterances such as "okay" [5, 6]. For this reason, the segments that we score are web pages and discussion site messages, rather than sentences.

Among colloquial Finnish, the collected corpora contained literary Finnish, foreign language, and even garbage such as HTML code that had slipped through the preprocessing scripts. The following scoring methods were targeted to separate such noise from the relevant text segments.

- **avg-unigram-count**. Word unigram counts are calculated from the entire text data. The score of a text segment is the average of the counts of the words in the segment.
- **median-unigram-count**. The score of a text segment is the median of the counts of the words in the segment. The reasoning is that garbage segments often contain short words that by chance are very common in Finnish language, and increase average unigram count.
- **devel-lp-ngram**. An n-gram model is estimated from the entire training data, and with a segment removed. The decrease in development data log probability when a segment is removed, is the score of the segment. This is the selection criterion used by Klakow [4].
- **devel-lp-ngram-topic**. As *devel-lp-ngram*, but filtering is applied per discussion site conversation instead of per discussion site message. Longer text segments allow more reliable probability estimates, but less fine-grained filtering.

4.4. Finding optimal filtering threshold

After every text segment has been assigned a score, those that have a score below a filtering threshold, will be excluded from the training data. We optimize a different threshold for every corpus, using the following method: The segments are sorted from the highest scoring to the lowest scoring. The training set is grown by gradually including more and more text, starting from the highest scoring segment. A bigram morph model is estimated from the training set, and development set perplexity is computed, at frequent intervals. Then we find the threshold score that minimizes perplexity.

It would be computationally too expensive to resegment the vocabulary into morphs every time training data is increased. We found it adequate to segment each corpus once, even though this means that with less training data, not all the morphs necessarily occur in the language model. The number of OOV morphs is significant only with very little training

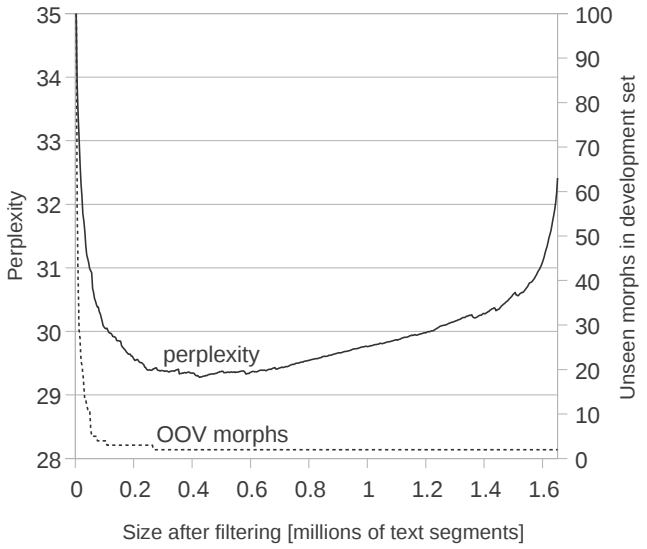

Figure 3: Perplexity and the number of OOV morphs on a held-out data set, with a growing amount of training data included in the order of *devel-lp-1gram* score

data. Figure 3 shows how perplexity and OOV rate behave as a function of included training data, with a fixed morph segmentation.

5. Experimental setup for language model evaluation

5.1. Speech recognizer

The speech recognition experiments were carried out using Aalto ASR system [15]. Our baseline model for recognizing standard Finnish has been trained on planned speech from the SPEECON [11] corpus. The model used in these experiments was trained on the SPEECON data, augmented with 176 minutes of our new development data, and 622 minutes of audio from FinDialogue, the conversational part of the FinINTAS corpus [16].

5.2. Error measure

Phonetic variation also creates challenges when measuring recognition accuracy. As most of the words can be pronounced in several slightly different ways, and the words are written out as they are pronounced, it would be harsh to compare recognition against the verbatim phonetic transcription. Thus word forms that are simply phonetic variation were added as alternatives in the reference transcriptions. This caused a large amount of manual work in top of transcription, since the added alternative pronunciations depend also on the meaning of the word, i.e. the context needs to be considered when adding alternations.

It has been customary in Finnish language speech recognitions to use letter error rate (LER) as the measure of speech recognition accuracy. We are not yet sure how to implement LER in the presence of a large number of alternative hypotheses of varying length, so this paper uses word error rate (WER).

5.3. Language models

Simply concatenating the data sets to estimate a language model would result in a model that is dominated by the biggest corpora, and performs poorly. A popular approach to combining different corpora is by linear interpolation of the language model probabilities. With many corpora, this becomes inefficient, requiring the decoder to evaluate every model for each possible word expansion. We used an approximative approach, where the probabilities of all observed n-grams are obtained by interpolating component model probabilities, and the remaining probabilities are computed to normalize the model [17]. The component weights are computed by optimizing development data (DEVEL1 + DEVEL2) perplexity. All the language models used in these experiments were pruned by removing n-grams whose removal caused less than 5×10^{-10} increase in training data perplexity.

We wanted to eliminate the effect of vocabulary selection from the data selection experiments, so all the word models were trained with the same 87,971 word vocabulary consisting of the words that occur at least 40 times in data sets WEB1 to WEB5. This left 8.4 % of word tokens in the verbatim evaluation set transcriptions out of vocabulary. However, since the reference transcriptions include alternative word forms, the recognizer may occasionally recognize a word correctly, even if the exact word form is not included in the vocabulary. Taking the alternatives into account, 6.0 % of the evaluation set word tokens could not be recognized with this vocabulary.

6. Results

6.1. Filtering evaluation

Table 2 shows the total size of data sets WEB1 to WEB5 before and after filtering, and error rates given by 4-gram language models on the evaluation data. *devel-lp-1gram* was the most effective filtering method. It resulted in a small data set (23 % of the original word tokens), and 1.0 % reduction in WER. In line with Klakow's results [4] filtering worked slightly better with unigram than bigram log probability.

avg-unigram-count filtering did not improve error rates, on contrary to *median-unigram-count*. Performing filtering only on conversations was too coarse-grained. *devel-lp-1gram-topic* reduced the amount of text and recognition errors minimally.

6.2. Comparison against existing corpora

Our current baseline language models have been created using 143 million words from the Finnish Language Text Collection (FTC), an electronic collection of Finnish text from newspa-

Filtering algorithm	WEB1	WEB2	WEB3	WEB4	WEB5	Interp.	Words
unfiltered	63.6	66.4	65.9	60.4	60.6	59.2	73,787,371
avg-unigram-count	63.5	66.5	65.8	59.8	60.9	59.6	35,426,285
median-unigram-count	63.5	66.1	65.9	59.6	59.9	58.6	37,637,867
devel-lp-1gram	63.4	**65.2**	65.5	59.5	**58.5**	**57.5**	**16,936,104**
devel-lp-2gram	63.5	65.7	**65.3**	**58.9**	59.1	57.7	19,059,831
devel-lp-1gram-topic	**63.3**	66.0	65.6	60.4	60.3	59.1	69,710,151
devel-lp-2gram-topic	63.5	65.7	**65.3**	60.4	60.3	59.5	69,708,077

Table 2: Recognition results from language models trained on filtered and unfiltered web data sets and an interpolated language model, and remaining total training data sizes in words

pers, journals, and books from the 1990's. Word error rates around 10 % on literary Finnish can be achieved with language models estimated from this corpus alone. Recently we have acquired two new corpora: 442,000 word "Helsingin puhekielen korpus" (HPK), a collection of interviews in dialectal language from the 1970's [18], and FinDialogue (FD), 81,000 words of conversational Finnish from FinINTAS corpus [16]. We have evaluated the web data in a speech recognition experiment against these corpora. All these corpora are either available or becoming available from CSC—IT Center for Science in Finland.

The comparison in Table 3 shows how poorly the existing corpora match colloquial Finnish speech. The collected web data alone performs better than the previous corpora combined with interpolation. WEBfilt is the web data after *devel-lp-1gram* filtering. It outperforms the previous corpora by 3.8 % in terms of word error rate. When the web data is combined with the previous corpora, WER is reduced by 7.0 %. This is a clear improvement in performance, given the amount of evaluation data, 44 minutes of speech from 17 speakers.

While filtering improved WER significantly when using only web data, when interpolating with the other corpora, it reduced model size, but did not improve WER. This result suggests that the interpolation may not be optimal. It might be beneficial to filter also the literary Finnish corpora, or try different adaptation techniques.

Training set	N-grams	WER	PPL
FTC	20,780,423	72.2	6364
FTC+HPK+FD	8,772,995	59.8	674
WEB	15,803,759	59.2	652
WEBfilt	3,694,060	57.5	589
FTC+HPK+FD+WEB	14,884,046	**55.6**	**493**
FTC+HPK+FD+WEBfilt	5,429,240	55.7	496

Table 3: Language model sizes, recognition results, and perplexities from models interpolated from existing corpora and the collected web data

By combining the web data with existing corpora, we obtained 55.6 % WER. This can be compared to 61.9 % WER we obtained with acoustic model trained only on SPEECON corpus, using the same language model. The improvement is significant, although we still have only little colloquial Finnish speech data for acoustic model training.

Perplexity can be used to evaluate how well a language model alone performs on colloquial Finnish text. The perplexities in Table 3 were computed on the verbatim transcripts of the evaluation data, i.e. considering only the exact word forms as they were pronounced. They show even greater improvement than the speech recognition experiments, in how well the data sets match the evaluation data, indicating that the poor recognition results may partly be due to the acoustic model trained on mostly literary Finnish matching poorly with conversational speech.

For comparison, we have included some results from morph models in Table 4, although so far we have not been able to get morph-based recognition on par with word-based recognition on colloquial Finnish. The morph results are from interpolated 5-gram morph models. Morph segmentations were computed using Morfessor Baseline (MDL) algorithm [13] from words that occur at least three times in the training data, with equal weight on each word. Resulting morph vocabularies ranged from 107,000 to 130,000 morphs.

Training set	N-grams	WER
FTC+HPK+FD	11,374,836	63.9
FTC+HPK+FD+WEB	10,558,474	**58.8**
FTC+HPK+FD+WEBfilt	5,385,316	59.4

Table 4: Language model sizes and recognition results from morph-based models interpolated from existing corpora and the collected web data

There was no clear difference in the recognition results between the radio conversations and the student conversations. The podcast conversations gave highest error rates, presumably because they contain some uncommon technological jargon.

7. Conclusions

We have collected large amounts of language model training material for colloquial Finnish from the Internet, and pruned it effectively, reducing the data size, language model perplexity, and speech recognition error rates, with very limited development data available. We have also described why the unlimited nature of the vocabulary and pronunciation variation

makes this task, as well as speech recognition, particularly difficult on colloquial Finnish. The standard approach to unlimited vocabulary in Finnish is language models based on subword units. We have found morph-based language models useful in filtering text of a highly agglutinative language, but so far traditional word-based language models have worked best for the recognition task, at least in terms of word error rate. We hope the new cleaned-up data sets will help us collect even more data and address modeling the lexicon and the pronunciation variation of colloquial Finnish in our future research to develop effective statistical models for ASR and MT.

8. Acknowledgement

This work was financially supported by the Academy of Finland under the grant number 251170 (Finnish Centre of Excellence Program (2012–2017)) and Mobster project funded by Tekes.

9. References

[1] Seppo Enarvi, *Finnish Language Speech Recognition for Dental Health Care*, Licentiate thesis, Aalto University School of Science, Espoo, Finland, Mar. 2012.

[2] L. M. Määttänen, "Puheenomaisten piirteiden ilmeneminen erityyppisissä suomalaisissa kirjoitetuissa teksteissä," M.S. thesis, University of Groningen, Groningen, Netherlands, Aug. 2007.

[3] Pablo Fetter, Alfred Kaltenmeier, Thomas Kuhn, and Peter Regel-Brietzmann, "Improved modeling of oov words in spontaneous speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996)*, Washington, DC, USA, 1996, vol. 1, pp. 534–537, IEEE Computer Society.

[4] Dietrich Klakow, "Selecting articles from the language model training corpus," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*. 2000, vol. 3, pp. 1695–1698, IEEE Computer Society.

[5] Abhinav Sethy, Panayiotis G. Georgiou, and Shrikanth Narayanan, "Text data acquisition for domain-specific language models," in *Proc. 2006 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2006, EMNLP '06, pp. 382–389, Association for Computational Linguistics.

[6] Abhinav Sethy, Panayiotis G. Georgiou, Bhuvana Ramabhadran, and Shrikanth S. Narayanan, "An iterative relative entropy minimization-based data selection approach for n-gram model adaptation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 13–23, 2009.

[7] Robert C. Moore and William Lewis, "Intelligent selection of language model training data," in *Proc. ACL 2010 Conference Short Papers*, Stroudsburg, PA, USA, 2010, ACLShort '10, pp. 220–224, Association for Computational Linguistics.

[8] Milind Mahajan, Doug Beeferman, and X.D. Huang, "Improved topic-dependent language modeling using information retrieval techniques," in *Proc. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999)*. 1999, vol. 1, pp. 541–544, IEEE Computer Society.

[9] Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 5, no. 1, pp. 3:1–3:29, Dec. 2007.

[10] Ivan Bulyko, Mari Ostendorf, Manhung Siu, Tim Ng, Andreas Stolcke, and Özgür Çetin, "Web resources for language modeling in conversational speech recognition," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 5, no. 1, pp. 1:1–1:25, Dec. 2007.

[11] Dorota J. Iskra, Beate Grosskopf, Krzysztof Marasek, Henk van den Heuvel, Frank Diehl, and Andreas Kießling, "SPEECON - speech databases for consumer devices: Database specification and validation," in *Proc. Third International Conference on Language Resources and Evaluation (LREC 2002)*, Canary Islands, Spain, May 2002.

[12] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, Oct. 2006.

[13] Mathias Creutz and Krista Lagus, "Unsupervised discovery of morphemes," in *Proc. ACL 2002 workshop on morphological and phonological learning*, Stroudsburg, PA, USA, 2002, vol. 6 of *MPL '02*, pp. 21–30, Association for Computational Linguistics.

[14] Jeff A. Bilmes and Katrin Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT/NAACL-2003)*, Edmonton, Alberta, May/June 2003, vol. 2 of *NAACL 2003–short papers*, pp. 4–6.

[15] Teemu Hirsimäki, Janne Pylkkönen, and Mikko Kurimo, "Importance of high-order n-gram models in

morph-based speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 724–732, 2009.

[16] Mietta Lennes, "Segmental features in spontaneous and read-aloud Finnish," in *Phonetics of Russian and Finnish. General Introduction. Spontaneous and Read-Aloud Speech*, Viola de Silva and Riikka Ullakonoja, Eds., pp. 145–166. Peter Lang GmbH, 2009.

[17] Andreas Stolcke, "SRILM—an extensible language modeling toolkit," in *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 901–904.

[18] Heikki Paunonen, *Suomen kieli Helsingissä: huomioita Helsingin puhekielen historiallisesta taustasta ja nyky-variaatiosta*, Helsinki: Helsingin yliopiston suomen kielen laitos, 1995.