



Combination of Probabilistic and Possibilistic Language Models

Stanislas Oger, Vladimir Popescu, Georges Linarès

Laboratoire d'Informatique d'Avignon (LIA)
University of Avignon, France

{stanislas.oger, vladimir.popescu, georges.linares}@univ-avignon.fr

Abstract

In a previous paper we proposed Web-based language models relying on the possibility theory. These models explicitly represent the possibility of word sequences. In this paper we propose to find the best way of combining this kind of model with classical probabilistic models, in the context of automatic speech recognition. We propose several combination approaches, depending on the nature of the combined models. With respect to the baseline, the best combination provides an absolute word error rate reduction of about 1% on broadcast news transcription, and of 3.5% on domain-specific multimedia document transcription.

Index Terms: language models, world wide web, possibility measure, automatic speech recognition

1. Introduction

The huge amount of textual data that is accessible *via* Internet motivated many research efforts in the last decade, particularly in the fields of speech and natural language processing. For automatic speech recognition (ASR), the most developed approach consists in extracting a subset of the indexed text data and in estimating language models with this text [1, 2, 3]. This approach demonstrated its efficiency on low-resourced tasks, where ASR systems suffer from a lack of large text corpora, for example on under-resourced languages or on domain-specific tasks. In other cases, Web-based language models (LM) are usually less accurate than corpus-based n -gram models. The reason for this disappointing conclusion is that the word sequence statistics on the Web result from very different contexts, according to the source, the topic, the editorial style, etc. Training a good LM on such highly heterogeneous data usually leads to very generic LMs that finally do not sufficiently match the usage conditions. Although the n -gram distributions obtained on the Web are poorly-accurate for a specific application context, the Web offers a close-to-full coverage of most of the spoken languages, and a word sequence being unreachable through a search engine usually means that it is an impossible n -gram.

Unfortunately, the integration of such information in a classical LM is difficult. LMs have to be estimated on corpora that cover only a part of the target language, and unseen event probabilities are interpolated from lower-order n -grams. This back-off strategy is essential, even though it potentially leads to over-scoring of impossible word sequences.

In [4], we proposed a new paradigm for language modeling where this kind of information can be exploited. The resulting models rely on measures derived from the possibility theory [5]. The basic idea of this theory is that probability measures indiscriminately represent two concepts that are fundamentally

different: the *uncertainty* and the *imprecision*. For example, the probability of a hypothesis such “she is 23 years old” indicates that she may be around 23 years old, but also how much we are confident in this hypothesis. The possibility theory offers a theoretical framework for explicitly modeling the possibilities of hypothetical events, thus being an alternative or a complement to the probability theory. The use of such a possibility measure for language modeling encounters two major problems. The first one is how to estimate the possibility measure empirically, especially on the Internet. The second problem is how to provide the best integration of the possibilistic scores in speech recognition systems, which are mainly based on Bayesian statistics. The first problem has been tackled in [4], and the second is the focus of this paper. In this perspective, we propose three ways of combining the probabilities and possibilities: one relying on the possibility distribution seen as an upper bound of the probability distribution, one that consists in balancing the back-off coefficient of the probabilistic LM with the possibility, and the last is a linear combination.

In Section 2, we generalize the technique proposed in [4] for estimating possibility measures on the Web or closed corpora. In Section 3 we recall the technique used for estimating probabilities on closed corpora and the one proposed in [4] for estimating Web probabilities. Thus, we obtain four distinct LMs: Web- and corpus-based possibility and probability LMs. In Section 4 we propose three strategies for combining possibilistic and probabilistic LMs. The experimental results are analysed in Section 5. Finally, we discuss the interest of possibilistic models and propose some perspectives for future research.

2. Possibility measures

In [4], we proposed a Web-based possibility measure that relies on the existence or the non-existence of a word sequence on the Web. Here, we propose to generalize the formula for computing a possibility measure on any arbitrary corpus.

The web-based possibility measure proposed in [4] is based on the idea that the more long sub-sequences of a word sequence are numerous on the Web, the more possible the word sequence is. We can extend this concept to any corpus C .

First of all, for each desired language model order n , we recursively construct a distinct set of possibility distributions π_n^c to π_1^c , according to Equation 1:

$$\pi_n^c(W) = \frac{|W_n \cap C_n| + \alpha \cdot |W_n \setminus C_n| \cdot \pi_{n-1}(W)}{|W_n|} \quad (1)$$

where W is a sequence of n or more words, W_n is the set of word sequences of size n in W , C_n is the set of word sequences of size n in the corpus, and $0 \leq \alpha \leq 1$ is the back-off coefficient. The terminal condition for the recursion is $\pi_0^c(W) = 0$.

This research has been funded by the National Research Authority (ANR), AVISON project (ANR-007-014).

The higher the number of sequences W present on the corpus is, the higher the possibility of W is.

For a given word sequence, this distribution expresses the number of its sub-sequences of length n that are present on the Web, with respect to the total number of its sub-sequences of length n . The possibility mass that is lost because of the absence of sub-sequences of length n on the corpus, is redistributed to the possibility measure of lower order.

The possibility distribution π_n^c defined above, in Equation 1, allows us to derive the possibility measure Π_n^c , according to Equation 2:

$$\Pi_n^c(A) = \max_{W \in A} (\pi_n^c(W)) \quad (2)$$

where A is a set of sequences of n or more words.

We have thus proposed a formula that allows us to estimate a possibilistic distribution on the Web, as well as on a classical text corpus. Unlike in classical n -gram-based language models, the possibility measures are estimated directly on entire hypotheses, without computing individual possibility measures for sub-hypotheses.

3. Web-based probability measures

As for the possibilistic measure, we propose a method for estimating probability LMs on any type of corpus, as well as on the Web. For textual corpora, we can estimate classical n -gram distributions with a state-of-the-art back-off technique, such as the modified Kneser-Ney method [6]. However, as far as the Web is concerned, we cannot compute probabilistic distributions based on n -gram frequencies, since this information is not available. In [4], we proposed a Web-based n -gram probability measure relying on counts of documents that contained a given n -gram. For a given word w_i , we first denote by ψ_i^n its history of size $n - 1$ in an n -gram: $\psi_i^n = w_{i-n+1} \dots w_{i-1}$. Thus, in order to obtain the probability of a Web n -gram, we use Equation 3:

$$P_{\text{web}}(w_i|\psi_i^n) = \frac{H(\psi_i^n, w_i)}{H(\psi_i^n)} \quad (3)$$

where $H(W)$ is the number of documents that contain the word sequence W retrieved by the search engine, and n is the order of the model. In order to cope with the zero probabilities, we interpolated our distribution with lower-order distributions. Probabilities are therefore computed by using Equation 4:

$$P_{\text{web}}^*(w_i|\psi_i^n) = \alpha_1 \cdot P_{\text{web}}(w_i|\psi_i^n) + \alpha_2 \cdot P_{\text{web}}(w_i|\psi_i^{n-1}) + \dots + \alpha_n \cdot P_{\text{web}}(w_i) \quad (4)$$

where α_i are positive real numbers such that $\sum_{i=1}^n \alpha_i = 1$.

Therefore, we have proposed a way of computing a probabilistic n -gram distribution on the Web, and we already have probabilistic distributions for textual corpora.

4. Hybrid language models

In previous sections, we have proposed possibilistic and probabilistic measures that can be computed on the Web, as well as on classical textual corpora. We propose here several ways of combining these models.

4.1. Improved back-off models

The n -gram LM back-off are poorly estimated in general, even *via* the Kneser-Ney method, and we think that combining it with the proposed measures can be beneficial. We proposed

two equations, depending on the nature of the measure to be combined. If the measure is a probability, Equation 5 applies:

$$\hat{P}(w_i|\psi_i^n) = \begin{cases} \alpha \cdot P_{\text{LM}}(w_i|\psi_i^n) + (1 - \alpha) \cdot P_{\text{web}}^*(w_i|\psi_i^n), & \text{if } w_i \in U_{\psi_i^n} \\ \beta \cdot P_{\text{LM}}(w_i|\psi_i^n), & \text{otherwise} \end{cases} \quad (5)$$

where α is a positive, empirically chosen, weighting factor, $U_{\psi_i^n}$ is the set of words w_i with the history ψ_i^n of size $n - 1$, for which the baseline LM has to backoff, and β is a normalization factor, defined according to Equation 6:

$$\beta = \frac{1 - \sum_{u \in U_{\psi_i^n}} \hat{P}(u|\psi_i^n)}{1 - \sum_{u \in U_{\psi_i^n}} P_{\text{LM}}(u|\psi_i^n)} \quad (6)$$

Otherwise, if the measure is a possibility, Equation 7 applies:

$$\hat{P}(w_i|\psi_i^n) = \begin{cases} \Pi_n(\{\psi_i^n, w_i\}) \cdot \alpha(\psi_i^n) \cdot P(w_i|\psi_i^{n-1}), & \text{if } w_i \in U_{\psi_i^n} \\ \beta \cdot P_{\text{LM}}(w_i|\psi_i^n), & \text{otherwise} \end{cases} \quad (7)$$

where $\alpha(\psi_i^n)$ is the baseline LM back-off factor. We thus redistribute, through the β factor defined in Equation 6, the probability mass wrongly assigned to impossible events according to the Web, to the events that were seen in the training corpus.

4.2. Possibilities as upper bounds of probabilities

There are several definitions of the relation between possibilities and probabilities. Here, we use the definition provided in [7], where it is stated that, for a probability measure P defined on Ω , the possibility measure Π that corresponds to P satisfies Equation 8:

$$\forall A \subseteq \Omega, P(A) \leq \Pi(A) \quad (8)$$

We can further use this property for improving a probabilistic LM. It is likely that the probability assigned to events by the general LM be sometimes greater than the possibility assigned to this event by the possibilistic LM. Hence, we can redistribute this exceeding probability mass between the well-learned events in the general LM. Equation 9 formalizes this idea:

$$\hat{P}(w_i|\psi_i^n) = \begin{cases} \Pi_n(\psi_i^n, w_i)^\alpha, & \text{if } \Pi_n(\psi_i^n, w_i)^\alpha < P_{\text{LM}}(w_i|\psi_i^n) \\ \beta \cdot P_{\text{LM}}(w_i|\psi_i^n), & \text{otherwise} \end{cases} \quad (9)$$

where Π_n is a possibility measure, α is a scaling factor for controlling the fraction of the probabilities affected by the cut; with $\alpha = 0$, no probabilities are modified. β is a normalization factor, defined similarly to Equation 6.

Starting from this idea, the Web- and corpus-based possibilities can be seen as upper bounds of Web- and corpus-based probabilities.

4.3. Log-linear combination

The probabilistic and possibilistic LMs can also be considered as complementary linguistic scores. In a typical ASR system, each hypothesis is assigned a score, computed as a log-linear combination between an acoustic score ($P(X|W)$), and a weighted linguistic probability ($P(W)^\alpha$, with $0 \leq \alpha \leq 1$). For improving this score, we could add other linguistic information to this score, by adding terms to the log-linear combination.

For instance, linguistic possibility measures can be integrated into the ASR framework by replacing the classical best-hypothesis search equation by Formula 10:

$$\hat{W} = \arg \max_W P(X|W) \times P(W)^\alpha \times \Pi(W)^\beta \quad (10)$$

where W is a word hypothesis, X is the sequence of acoustic observations, $P(X|W)$ is the acoustic score, $P(W)$ is the linguistic probability, $\Pi(W)$ is the linguistic possibility, α and β are positive, empirically chosen, weighting factors.

According to this approach, we can combine in all possible ways the four measures that we proposed for estimating the global score of the hypotheses: Web- and corpus-based possibility measures, and Web- and corpus-based probability measures.

5. Results

In this section, we evaluate the proposed models and their combinations on two ASR tasks: broadcast news transcription, and domain-specific spoken discourse transcription.

5.1. Experimental setup

For assessing the proposed methods in the two transcription tasks, we used the Laboratoire Informatique d'Avignon (LIA) broadcast news transcription system, SPEERAL [8]. This system is an A* decoder based on state-dependent hidden Markov models for acoustic modeling, and on a n -gram LM.

For the broadcast news transcription task, we used the test corpus of the HUB4'98 campaign [9], of about 3 hours of English broadcast news. The baseline LM is a 65k word classical 3-gram, estimated on 2.7G words from the Gigaword, North American News and HUB4 corpora, by using the modified Kneser-Ney smoothing technique. The transcription word error rate (WER) of the test corpus with this configuration, without speaker adaptation, is 27.0%.

For the domain-specific transcription task, we used 4 hours from the English AVISON corpus, which contains recorded surgery-related discourse. A combined 65k word LM is used, by interpolating general 3-grams learned on the HUB4 English corpus, with 3-grams estimated on all the reference transcriptions available in the AVISON training corpus, by relying, here as well, on the modified Kneser-Ney smoothing technique. A baseline WER of 27.9%, without speaker adaptation, was obtained on this domain-specific corpus.

The direct use of the proposed Web-based LMs in the search algorithm of the ASR system would lead us to submit too many queries to the Web search engines. This is why a 100-best decoding is done instead, with the baseline n -gram LM, which produces the top 100 recognition hypotheses; the proposed Web-based LMs are used for rescoreing these hypotheses in combination with the acoustic score. The Google search engine is used for processing Web queries.

The optimization of the weight factors for the log-linear combinations and the back-off coefficients is performed in the K -fold cross-validation framework, with $K = 10$: firstly, the test corpus is partitioned into ten sub-corpora; then, the coefficients are optimized on nine partitions and tested on the tenth. This last step is repeated ten times, with each of the ten sub-corpora used exactly once as test data. The global error rate is the number of erroneous words of all the test partitions over the number of words in the test corpus.

5.2. Individual possibilistic and probabilistic measures

The proposed Web-based measures perform better with high n -gram orders. In order to compare these measures with the baseline LM, we estimated corpus-based n -gram models of orders

from 4 to 6, on the same data as the 3-gram models. Only the results obtained for 3- and 6-gram models are shown, because they are the most representative. The row P_c in Table 1 contains the WER obtained by the baseline corpus-based n -gram models of orders 3 and 6 on the 100-best rescoreing task.

The rows P_w , Π_c and Π_w in Table 1 contain respectively the results of the Web-based probabilistic LM, the corpus-based possibilistic LM, and the Web-based possibilistic LM alone.

The high-order Web-based LMs perform well. This is explained by the huge size of this corpus, that allows us to estimate accurate high-order possibilities and probabilities.

The performances of the corpus-based possibility model are worse than all the other models. We believe that this happens because textual corpora are too small for estimating reliable possibility measures. The non-existence of a given n -gram in a corpus does not indicate that it is impossible.

The Web-based probability model works well on the two corpora, which indicates that the proposed document-frequency estimation is relevant for taking advantage of the information present on the Web.

The Web-based possibility model performs well on the domain-specific AVISON corpus, whereas no improvement is obtained on the broadcast news corpus. This shows that this measure is relevant on the under-resourced domain covered by the AVISON corpus.

To conclude, the Web-based possibility measures are effective in the specialized domain, which shows that, as expected, they are an alternative to probability measures when available relevant corpora are not large enough.

5.3. Improved back-off models

The rows P_w BO P_c , Π_w BO P_c and Π_c BO P_c in Table 1 contain respectively the results of the Web-based probability used as corpus-based back-off coefficient, the Web-based possibility used as a corpus-based back-off coefficient, and the corpus-based possibility used as a corpus-based back-off coefficient.

All the modified back-off approaches improve the baseline corpus-based probabilities, which indicates that the corpus-based back-off probabilities become more accurate when they are combined with other kind of information.

Adding corpus-based possibility measures in the back-off slightly improves the performances. We obtain an absolute WER reduction of 0.6% for the AVISON corpus, and of 0.2% for HUB4. These results show that the possibility measures add information to the classical probabilistic modeling.

The results obtained with the Web-based probabilistic and possibilistic back-off are similar. However, the Web-based probabilistic back-off behaves slightly better: an absolute WER reduction of 1.9% is obtained for the AVISON corpus, whereas on HUB4 this reduction is of 0.3%, with respect to the corpus-based probability measure.

To conclude, the results show that the Web-based back-off is better than the corpus-based one, irrespective of the measures used.

5.4. Possibilities as upper bounds of probabilities

The rows $P_w \leq \Pi_c$, $P_c \leq \Pi_w$, and $P_w \leq \Pi_w$ in Table 1 contain respectively the results of the corpus-based possibilities used as upper bounds of Web-based probabilities, the Web-based possibilities used as upper bounds of corpus-based probabilities, and the Web-based possibilities used as upper bounds of Web-based probabilities.

The corpus-based possibilities do not bring any improvement when used as an upper bound for the Web-based probabilities. However, using the Web-based possibilities as an upper bound for the corpus-based probabilities yields an absolute WER improvement of 0.2% on HUB4, and of 2% on the AVISON corpus.

As expected, the best upper bound combination is obtained with the best probabilistic and possibilistic measures (both on the Web) and yields an absolute WER improvement of 0.8% for HUB4, and of 2.7% for the AVISON corpus, with respect to the corpus-based probability measure.

These results confirm that the corpus-based probabilistic LM assigns too high a probability mass to certain events, and that the Web-based possibilistic measure allows one to rescore these events.

5.5. Log-linear combination

Starting with the four measures that we proposed (corpus- and Web-based probabilities and possibilities), eleven combinations of two, three and four measures are possible. We present here the most interesting combinations.

The rows $P_w + P_c$, $P_w + \Pi_w$, and $P_c + \Pi_w$ in Table 1 contain respectively the results of the log-linear combination of the Web-based probabilities with the corpus-based probabilities, the Web-based probabilities with the Web-based possibilities, and the corpus-based probabilities with Web-based possibilities. The results are very different for HUB4 and for the AVISON corpus. On HUB4, none of the combinations is significantly better than the Web-based probabilities alone. On the AVISON corpus, the best combination, of the Web-based probability and possibility measures, allows for an absolute WER improvement of 3.3%. This result confirms that for the same ‘‘corpus’’ (here, the Web), the possibilistic measure adds information to the probabilistic measure.

The row $\Pi_w + \Pi_c + P_w + P_c$ in Table 1 contains the log-linear combination of the four proposed metrics: corpus- and Web-based probabilities and possibilities. The combination of the four measures yields, globally, the best performances. We thus obtain an absolute WER improvement of 3.5% on the AVISON corpus, which represents an improvement of 0.7% with respect to the best estimator alone (the Web-based possibility measure). For HUB4, the combination of the four measures yields an improvement of 0.6% on 3-gram LMs and 0.1% on 6-gram LMs, with respect to the best measure alone (the Web-based possibility measure). Hence, the log-linear combination of the four measures seems to be the best way of jointly using them.

6. Conclusion

In this paper we have generalized the formulation of an approach for measuring the possibility of a word sequence, by relying on the possibility theory. This extended formula allows us to estimate the measure on a classical closed corpus, or on a particular ‘‘corpus’’, the Web. Then, we have proposed several ways of combining this novel information source with classical probabilistic LMs in ASR systems. The resulting models have been evaluated on two ASR tasks. The first task boils down to transcribing broadcast news, which are well covered by currently available corpora, and hence, by well-trained classical LMs. The second task consists in transcribing specialized-domain data, where not enough available corpora are available for correctly training a classical LM.

By analyzing the results obtained, we can see that, first,

Table 1: WER [%] of the proposed LMs, depending on the order n of the LMs, on the HUB4 and AVISON corpora.

method	AVISON		HUB4	
	$n = 3$	$n = 6$	$n = 3$	$n = 6$
P_c	27.9	28.0	27.0	26.9
P_w	27.2	25.5	27.0	26.0
Π_w	28.5	25.2	27.7	26.9
Π_c	28.0	28.1	27.9	27.9
$P_w \text{ BO } P_c$	27.6	26.1	26.7	26.6
$\Pi_w \text{ BO } P_c$	27.5	26.5	26.8	26.7
$\Pi_c \text{ BO } P_c$	27.6	27.4	26.8	26.9
$P_w \leq \Pi_c$	27.8	25.6	27.2	26.1
$P_c \leq \Pi_w$	28.1	26.0	27.0	26.7
$P_w \leq \Pi_w$	27.6	25.3	27.0	26.1
$P_w + P_c$	27.1	25.4	26.6	25.9
$P_w + \Pi_w$	27.1	24.7	27.0	26.1
$P_c + \Pi_w$	27.7	24.8	26.8	26.6
$\Pi_w + \Pi_c$ $+ P_w + P_c$	26.8	24.5	26.4	25.9

the Web contains useful information for language modeling, even for broadcast news. The Web-based possibilistic measure alone performs better than the probabilistic measures alone for the domain-specific spoken discourse transcription. On broadcast news, the Web-based probabilistic model remains the best. In general, combining these linguistic measures improves the WER, compared to using each of the measures alone. Among the proposed combination methods, the log-linear combination provides the best results. Combining all the measures yields the best results on the two corpora and allows for an absolute WER decrease of 3.5% on AVISON and of 1% on HUB4 with respect to the baseline LMs.

These results show the advantages of combining the linguistic measures that we proposed. In the near future, we will study a way of integrating the linguistic measures directly in the decoder, in order to prune earlier the incorrect hypotheses.

7. References

- [1] C. Martins, A. Teixeira, and J. Neto, ‘‘Dynamic language modeling for a daily broadcast news transcription system,’’ in *Proceedings of the ASRU*, 2007, pp. 165–170.
- [2] M. Federico and N. Bertoldi, ‘‘Broadcast news LM adaptation over time,’’ *Computer Speech & Language*, vol. 18, no. 4, pp. 417–435, 2004.
- [3] A. Berger and R. Miller, ‘‘Just-in-time language modelling,’’ in *Proceedings of ICASSP*, 1998, vol. 2, pp. 705–708.
- [4] S. Oger, V. Popescu, and G. Linares, ‘‘Probabilistic and possibilistic language models based on the world wide web,’’ in *Proceedings of INTERSPEECH*, 2009, pp. 2699–2702.
- [5] D. Dubois, ‘‘Possibility theory and statistical reasoning,’’ *Computational Statistics and Data Analysis*, vol. 21, pp. 47–69, 2006.
- [6] J.T. Goodman, ‘‘A bit of progress in language modeling extended version,’’ Tech. Rep., Microsoft Research, 2006.
- [7] Didier Dubois and Henri Prade, *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, Plenum Press, 1988.
- [8] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, JF Bonastre, D. Massoné, and F. Béchet, ‘‘The LIA’s french broadcast news transcription system,’’ in *SWIM*, 2004.
- [9] R. Stern, ‘‘Specifications of the 1996 hub-4 broadcast news evaluation,’’ in *Proceedings of DARPA Speech Recognition Workshop*, 1997, pp. 7–14.