

Probabilistic and Possibilistic Language Models Based on the World Wide Web

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

Usually, language models are built either from a closed corpus, or by using World Wide Web retrieved documents, which are considered as a closed corpus themselves. In this paper we propose several other ways, more adapted to the nature of the Web, of using this resource for language modeling. We first start by improving an approach consisting in estimating n -gram probabilities from Web search engine statistics. Then, we propose a new way of considering the information extracted from the Web in a probabilistic framework. Then, we also propose to rely on Possibility Theory for effectively using this kind of information. We compare these two approaches on two automatic speech recognition tasks: (i) transcribing broadcast news data, and (ii) transcribing domain-specific data, concerning surgical operation film comments. We show that the two approaches are effective in different situations.

Index Terms: language modeling, World Wide Web, possibility measure, automatic speech recognition

1. Introduction

The quality of the n -gram language models (LM) depends on the size and quality of the corpus used for learning these models. Linguistic coverage cannot be exhaustive with any closed corpus, especially when particular domains are concerned.

The Web has been often used for estimating LMs. In most of the cases, this boils down to collecting domain-specific documents from the Web, and then estimating classical n -gram models on these documents [1], [2], [3]. A second marginal approach consists in estimating the n -gram probabilities by using directly statistics obtained through Web search engines [4], [5].

Several research reports show that, generally, n -grams estimated from the Web are less costly to obtain, but at the same time of a lower quality than LMs learned from closed corpora, mainly because the statistical distributions of the Web n -grams are not reliable [4]. Nevertheless, the Web is quite exhaustive, and the existence of a word sequence on the Web can constitute relevant information that should be integrated in the LMs. The integration of impossible sequences has been studied in [6], where the authors automatically produce impossible 2-grams and propose methods for integrating them into LMs. However, one of the main obstacles to this approach stems from the difficult *a priori* generation of the impossible 2-grams.

In this paper, we first present a framework for estimating n -gram probabilities starting from the statistics of a Web search engine, and we propose ways to combine these probabilities with classical LMs. Then, we reconsider the information obtained from the Web: rather than approximating n -gram probabilities, we introduce a Web-based *possibilistic* measure [7]. We

propose several ways of combining classical LMs with information yielded by this measure. Finally, these strategies are compared on an automatic speech recognition (ASR) task, on two corpora: a multi-domain, news corpus (ESTER), and a domain-specific corpus (AVISON), of medical surgery recordings.

2. The Web as an Ad-hoc n -gram LM

N -gram probabilities are usually estimated from word sequence counting on corpora. Using the Web for computing these probabilities thus requires, for a given word sequence, the knowledge of its frequency in at least some of the documents on the Web. In order to do this, we can rely on statistics obtained with a Web search engine: most of them provide the number of documents that satisfy a given query; this query can be an n -gram word sequence. Using the number of documents that contain a specified word sequence, we can deduce the number of n -grams. This approach is presented in [5]. However, we can observe that, on the one hand, these authors do not present any result obtained by using directly the number of documents as an estimate of the number of n -grams. On the other hand, Zhu *et al.* use Web-based statistics only for backing off from classical, corpus-based LMs and do not use only Web-based probabilities as a linguistic score. It is precisely this last possibility that we explore here.

2.1. Estimating n -gram Probabilities from the Web

We propose to estimate the Web n -gram probabilities by relying directly on the number of documents that contain a given n -gram. For a given word w_i , we first note 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 1:

$$P_{\text{web}}(w_i|\psi_i^n) = \frac{H(\psi_i^n, w_i)}{H(\psi_i^n)} \quad (1)$$

where $H(S)$ is the number of documents that contain word sequence S retrieved by the search engine, and n is the order of the n -gram model. However, Equation 1 is not easy to use, because it assigns a zero probability to the word sequences that are not on the Web. To tackle this issue, we usually redistribute one part of the probability mass assigned to the events seen during training, to unseen events. Given that the necessary statistics for using a state-of-the-art backoff technique, such as the modified Kneser-Ney method [8], are not available by using the Web in this way, we will interpolate our distribution with lower order distributions, which were proven to work well [8]. Probabilities are therefore computed by using Equation 2:

$$P_{\text{web}}^*(w_i|\psi_i^n) = \alpha_1 \cdot P_{\text{web}}(w_i|\psi_i^n) + \alpha_2 \cdot P_{\text{web}}(w_i|\psi_i^{n-1}) + \dots + \alpha_n \cdot P_{\text{web}}(w_i) \quad (2)$$

where α_i are positive real numbers such that $\sum_{i=1}^n \alpha_i = 1$. However, this formulation exhibits a difficulty, related to the

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

estimation of the unigram probability.

In the Web context, the frequency of a word is computed as the number of Web documents that contain this word, and the size of the corpus corresponds to the total number of documents indexed by the search engine. For an estimation of the latter value, we use the number of documents that contain the most frequent word in the natural language of interest (for English, the word *the*), thus hoping to cover most of the documents, in the language of interest, that are indexed by the search engine. The unigram probability thus computed will never be zero if one makes sure that all the words in the vocabulary are present in at least one Web document.

We therefore propose a Web n -gram probability estimation method that does not result in zero probabilities, even for unseen word sequences. With Web n -gram probabilities thus defined, there are several manners of using them for computing word sequence probabilities.

2.2. The Web probabilities as a Better Backoff

A first approach, that has been recently used in [5], consists in using the Web probability in baseline LM backoff. This boils down to giving to the n -gram probabilities seen as such in the corpus a higher confidence level than to the Web-based probabilities. Let $U_{\psi_i^n}$ be the set of words w_i with the history ψ_i^n of size $n - 1$, for which the baseline LM has to backoff. In formal terms, this can be written as in Equation 3:

$$\hat{P}(w_i|\psi_i^n) = \begin{cases} \alpha \cdot P_{\text{LM}}(w_i|\psi_i^n) + \\ (1 - \alpha) \cdot P_{\text{web}}^*(w_i|\psi_i^n), & \text{if } w_i \in U_{\psi_i^n} \\ \beta \cdot P_{\text{LM}}(w_i|\psi_i^n), & \text{otherwise} \end{cases} \quad (3)$$

where α is a positive, empirically chosen, weighting factor, and β is a normalization factor, defined in Equation 4:

$$\beta = \frac{1 - \sum_{u \in U_{\psi_i^n}} \hat{P}(u|\psi_i^n)}{1 - \sum_{u \in U_{\psi_i^n}} P_{\text{LM}}(u|\psi_i^n)} \quad (4)$$

2.3. The Web Probabilities as a Complete Language Model

Another way of using the Web for building LMs is to consider that the probabilities estimated from the Web are reliable, and thus not to interpolate these probabilities with the LM learned from the corpus; this is shown in Equation 5:

$$\hat{P}(w_i|\psi_i^n) = P_{\text{web}}^*(w_i|\psi_i^n) \quad (5)$$

This approach is justified when the corpus used for learning the LM is too small or too poorly adapted to the task under purview. In Section 4.2 we present several experiments with these two approaches, in an ASR task.

3. The Web as a Possibility Estimator

Numerous research reports show how one can take advantage of the statistics of word sequences on the Web. In the previous section, we have just proposed another such manner. Nonetheless, the *absence* of an n -gram from the Web could represent relevant information, which could be integrated in an LM. To our knowledge, this information has never been studied in the literature. Possibility theory [7] provides a theoretical framework for modeling this information.

3.1. Background on Possibility Theory

Possibility theory is a mathematical framework devoted to handling uncertainty resulted from incomplete knowledge [7].

Therefore, it complements Probability theory. Originally designed in order to formalize the notion of linguistic uncertainty [7], Possibility theory has been recently given a formal account akin to Probability theory, by relying on measure-theoretic concepts [9], thus transforming it into a quantitative framework for reasoning with incomplete knowledge.

A first important notion is that of *possibility distribution*, which is a mapping π from a set E of events to the unit interval $[0; 1]$. This function represents the knowledge distinguishing what is plausible from what is less plausible, what is atypical from what is “normal”. The following conventions apply to function π : (i) if $\pi(e) = 0$, then event e is rejected as impossible; (ii) if $\pi(e) = 1$, then event e is totally possible (plausible).

In a manner akin to Probability theory, a *possibility measure* can be computed from the possibility distribution, if the set of events is bounded [9]. Formally, we consider a set E of events. A possibility measure Π can be defined on the set of events E as $\Pi(E) = \max_{e \in E} \pi(e)$. In other words, $\Pi(E)$ evaluates to what extent the set E of events is consistent with the knowledge π . For any two subsets A and B of E , the joint possibility measure of A and B is constrained by:

$$\Pi(A \cap B) \leq \min(\Pi(A), \Pi(B)) \quad (6)$$

3.2. A Web-based Possibility Measure

In this section, we show how a possibility measure can be obtained for word sequences, by using statistics from the Web.

The possibility measure has to represent the possibility that a word sequence exists. For this, we rely on the existence of this sequence and of its sub-sequences on the Web. By *existence* on the Web, we mean here the fact that there exists at least one Web document that contains the word sequence under discussion. The idea is that the more long sub-sequences of the word sequence exist on the Web, the more the word sequence is possible. However, one needs to limit the search of sub-sequences in order to obtain a reliable measure. Indeed, the smaller the corpus considered for computing the possibilistic measure is (here, the Web), the less the non-existence of long sequences is significant.

First of all, for each desired LM order n , we recursively construct a distinct set of possibility distributions π_n to π_1 , according to Equation 7:

$$\pi_n(W) = \frac{|W_n \cap \text{Web}_n| + \alpha \cdot |W_n \setminus \text{Web}_n| \cdot \pi_{n-1}(W)}{|W_n|} \quad (7)$$

where W is a sequence of n or more words, W_n is the set of word sequences of size n in W , Web_n is the set of word sequences of size n on the Web, \setminus is the set subtraction operator and $0 \leq \alpha \leq 1$ is the backoff coefficient. The terminal condition for the recursion is $\pi_0(W) = 0$.

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 Web, is redistributed to the possibility measure of lower order.

The set of possibility distributions previously defined allows us to construct a corresponding set of possibility measures Π_n , according to Equation 8:

$$\Pi_n(\Theta) = \max_{W \in \Theta} (\pi_n(W)) \quad (8)$$

where Θ is a set of sequences of n or more words; if Θ has only one element W , then $\Pi_n(\{W\}) = \pi_n(W)$.

3.3. Possibility as a Backoff Modifier

The possibility measure previously introduced informs us on the confidence that we can have in the existence of a word sequence. If we have a higher confidence in the training corpus than in the Web, then all the n -grams seen in this corpus are totally possible ($\pi_n(\psi_i^n, w_i) = 1$). On the contrary, the n -grams composed by using backoff strategies are subject to controversy. We thus propose to weigh the probability that the LM assigns to the unseen n -grams in the training corpus, with the possibility estimated from the Web. This idea is formalized in Equation 9:

$$\hat{P}(w_i|\psi_i^n) = \begin{cases} \Pi_n(\{\psi_i^n, w_i\}) \cdot \alpha(\psi_i^n) \cdot P(w_i|\psi_i^{n-1}), \\ \text{if } w_i \in U_{\psi_i^n} \\ \beta \cdot P_{\text{LM}}(w_i|\psi_i^n), \text{ otherwise} \end{cases} \quad (9)$$

where $\alpha(\psi_i^n)$ is the baseline LM backoff factor. We thus redistribute, through the β factor defined in Equation 4, the probability mass wrongly assigned to impossible events according to the Web, to the events that were seen in the training corpus. The results of this approach are presented in Section 4.3.

3.4. Possibility Measure as a Stand-alone Metric

In the previously presented approach, the possibilities were seen as a backoff strategy when the LM was not “competent” enough *per se*. Nevertheless, possibilities can be also seen as a stand-alone linguistic measure and used as such, for example in combination with the acoustic score, in ASR tasks. Starting from Equation 7 where the Web-based possibility of a word sequence is defined, there are two ways of obtaining the possibility of a word sequence (i.e., a sentence). Such a sentence is denoted by S^m , the sequence of m words w_i , for $i \in \{1 \dots m\}$.

The first way boils down to considering each sentence as a set of word sub-sequences that can be assigned a possibility. Thus, S^m can be expressed as a set of n -sized word sequences S_n^m such that:

$$S_n^m = \{(\psi_i^n, w_i), \text{ for } i \in \{n \dots m\}\} \quad (10)$$

When we do not have complete information on an event, Possibility Theory compels us to choose the maximal estimate for the value of the possibility measure for this event. Thus, we can use the equality case in Inequality 6, for assigning a possibility to S_n^m (given by Equation 10):

$$\begin{aligned} \Pi_n(S_n^m) &= \min(\Pi_n(\{\psi_n^n, w_n\}), \Pi_n(S_n^m \setminus \{\psi_n^n, w_n\})) \\ &= \min(\Pi_n(\{\psi_n^n, w_n\}) \dots \Pi_n(\{\psi_m^n, w_m\})) \\ &= \min(\pi_n(\psi_n^n, w_n) \dots \pi_n(\psi_m^n, w_m)) \end{aligned} \quad (11)$$

The shortcoming of this first approach is that it reduces the possibility of a hypothesis to its least possible element; thus, this approach is not representative enough. Therefore, we propose a second way of representing the possibility of the word sequence S^m of size m greater than n ; this consists in directly applying Equation 7 on S^m :

$$\Pi_n(S^m) = \pi_n(S^m) \quad (12)$$

This last equation leads us to assign a possibility that is more representative of the whole sentence S^m .

The results of these approaches are presented in Section 4.4.

4. Experimental Results

The different approaches proposed in this paper for the use of Web information in LM building are evaluated on two typical ASR scenarios: (i) transcribing French broadcast news, with a baseline LM trained on a well-targeted corpus, and (ii) a domain-specific English discourse transcription task, with a baseline LM trained with a few data, the only available.

4.1. Experimental Configuration

For assessing the proposed methods in the two transcription tasks, we used the LIA broadcast news transcription system, SPEERAL [10]. This system is an A* decoder based on state-dependent hidden Markov models for acoustic modeling, and on a 3-gram LM.

For the broadcast news transcription task, we used about 6 hours from the test corpus of the ESTER 2005 campaign. The baseline LM is a 65k word classical 3-gram, estimated on 200M words from the French newspaper “Le Monde” and from the ESTER broadcast news training corpus (about 1M words), 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 24.4%.

For the domain-specific English transcription task, we used 2 hours from the English AVISON corpus, that 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, and by also relying on the modified Kneser-Ney smoothing technique. A baseline WER of 33.8%, without speaker adaptation, was obtained on this domain-specific corpus.

The direct use of the proposed linguistic Web estimators 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 3-gram LM, which produces the top 100 recognition hypotheses; the Web estimators are used for rescoring these hypotheses. This difficulty is tackled in the same way as in [5]. The Google search engine is used for processing Web queries.

4.2. Web-based n -gram Probability Results

In this subsection we present the results of the two strategies introduced in Section 2 for the use of the proposed Web probability estimation method.

The left part of Table 1, entitled “Web only”, contains the experimental results of the approach presented in Section 2.3, which consists in using only the Web probability model. The table contains the results with respect to the order of the Web LM, for the two corpora considered: ESTER and AVISON. Additional baseline classical LMs of orders $n \geq 3$ are learned, in order to compare them to high-order Web LMs. The results of these new baseline LMs on the task of rescoring 100-best hypotheses are reported in the right part of Table 1, entitled “Corpus-baseline”.

We observe that the 3-gram Web LM performs almost as well as the baseline LM on the AVISON corpus and a little worse on the ESTER corpus. However, by augmenting the order of the Web models, we observe a significant performance gain, whereas augmenting the order of the baseline models does not result in a comparable gain: the Web LM allows for a decrease of 2% absolute WER on the AVISON corpus and of 0.7% on the ESTER corpus, for $n = 6$, compared to the 6-gram baseline LMs. These results indicate that the Web statistics are not reliable for small n -gram orders ($n \leq 3$).

The center part of Table 1, entitled “Backoff Web”, contains the results of the second approach, described Section 2.2, which consists in using the Web n -grams as a backoff strategy, with a factor α that accords a weight of 0.9 to the Web n -grams.

As with the previous experiment, the results are better with high-order Web models; they confirm that the low order Web statistics are less reliable than the baseline backoff probabilities. However, with higher order Web LMs the performance in-

creases on both corpora. The best performance is a decrease of 2.2% in the absolute WER on the AVISON corpus with $n = 6$, with respect to the 6-gram baseline LM.

Table 1: AVISON and ESTER 100-Best rescoring WER [%] using various LMs, depending on the order n of the LMs.

n	Web only		Backoff Web		Corpus-baseline	
	ESTER	AVISON	ESTER	AVISON	ESTER	AVISON
3	24.8	33.7	24.7	33.7	24.4	33.8
4	23.7	32.8	24.2	32.0	24.2	33.5
5	23.6	32.5	24.2	31.3	24.2	33.3
6	23.5	31.3	24.1	31.1	24.2	33.3

4.3. The Possibility as a Backoff Modifier

The experimental results for the Web-based possibilistic approach to backoff, described in Section 3.3, on the ESTER and AVISON corpora, with respect to the order n of the Web possibility model, are reported in the left part of Table 2, entitled “Backoff-poss.”.

No gain is obtained on the ESTER corpus. However, the AVISON baseline LM is improved, even for the low-order models; this confirms the poor quality of the backoff probabilities of the AVISON baseline LM. This leads to an absolute WER gain of 1.7% on the AVISON corpus with $n = 3$.

4.4. The Possibility as a Standalone Metric

Here we present the experimental results for using the Web possibility as unique language rescoring metric; the theoretical details for this approach have been described in Section 3.4. The Web possibility scores are combined with the acoustic ones by using an empirically determined fudge factor.

The center part of Table 2, entitled “Min-poss.”, contains the experimental results of the first proposal, which consists in computing the hypothesis possibility as the minimum of their sub-sequence possibilities (Equation 11). The WER figures obtained on the ESTER and AVISON corpora are reported, depending on the order n of the Web possibility model.

A performance improvement is observed on both corpora: 0.6% of absolute WER reduction for AVISON with $n = 5$ and 0.3% for ESTER with $n = 3$, compared to the same order baseline models. However, this approach leads us to represent the possibility of a hypothesis by using the worst of the hypothesis sub-sequence possibilities, thus resulting in a loss of information.

The right part of Table 2, entitled “Global-poss.”, contains the experimental results of the second proposal, which consists in computing the Web possibility directly on the whole hypothesis, by using Equation 7. The WER figures obtained on the ESTER and AVISON corpora are reported, depending on the order n of the Web possibility model.

On the ESTER corpus, the absolute WER reduction with respect to the same order baseline model is of 0.5%, for $n = 5$, whereas by using the Web n -grams as described in Section 4.2, the WER reduction is of 0.7%. On the AVISON corpus, the absolute WER decrease is of 2.9% with $n = 6$; this is better than the WER reduction obtained with the Web n -grams. This last result shows that estimating Web probabilities is effective for improving well-trained LMs, whereas the Web possibilities are effective for improving poorly-trained LMs.

Table 2: WER [%] of the possibilistic LMs, depending on the order n of the possibility distribution, on the ESTER and AVISON corpora.

n	Backoff-poss.		Min-poss.		Global-poss.	
	ESTER	AVISON	ESTER	AVISON	ESTER	AVISON
3	24.5	32.1	24.1	33.9	24.1	31.8
4	24.3	31.9	24.2	33.1	23.8	31.5
5	24.3	31.9	24.2	32.7	23.7	30.6
6	24.3	31.7	24.2	33.3	23.9	30.4

5. Conclusion

We proposed two ways of using Web data as a source of linguistic information: (i) a probabilistic framework for using Web search engine document counts, and (ii) a possibilistic framework for using the presence of word sequences on the Web.

Our results show that the Web-based n -gram probability estimation methods proposed in this paper significantly outperform both Web-based backoff techniques and classical LM learning methods. Moreover, our results prove that the Web-based possibility measure used as unique linguistic score significantly improves poorly-trained LMs. The Web-based probabilistic approach allowed us to reduce the absolute ASR WER with 0.7% on the ESTER broadcast news corpus. The Web-based possibilistic approach allowed for an absolute ASR WER reduction of 2.9% on the AVISON domain-specific corpus. We have thus shown that Possibility Theory provides us with tools for effectively handling the information extracted from the Web in the context of ASR systems.

In the near future, we plan to experiment more effective ways of integrating these approaches in ASR systems, by acting directly on the word lattice for pruning wrong hypotheses earlier.

6. References

- [1] M. Federico and N. Bertoldi, “Broadcast news LM adaptation over time,” *Computer Speech & Language*, vol. 18, no. 4, pp. 417–435, 2004.
- [2] A. Berger and R. Miller, “Just-in-time language modelling,” in *Proc. ICASSP*, 1998, vol. 2, pp. 705–708.
- [3] I. Bulyko, M. Ostendorf, and A. Stolcke, “Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures,” in *Proc. HLT-NAACL*, 2003, vol. 2, pp. 7–9.
- [4] M. Lapata and F. Keller, “Web-based models for natural language processing,” *ACM Transactions on Speech and Language Processing*, vol. 2, pp. 1–31, 2005.
- [5] X. Zhu and R. Rosenfeld, “Improving trigram language modeling with the world wide web,” in *Proc. ICASSP*, 2001, vol. 1, pp. 533–536.
- [6] A. Brun, D. Langlois, and J.-P. Haton, “Improving statistical language models by removing impossible events,” in *Proc. SPECOM*, 2001.
- [7] D. Dubois, “Possibility theory and statistical reasoning,” *Computational Statistics and Data Analysis*, vol. 21, pp. 47–69, 2006.
- [8] J.T. Goodman, “A bit of progress in language modeling extended version,” Tech. Rep., Microsoft Research, 2006.
- [9] G. de Cooman, “Possibility theory I: the measure- and integral-theoretic groundwork,” *International Journal of General Systems*, vol. 25, pp. 291–323, 1997.
- [10] P. Nocéra, 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.