A new language model based on possibility theory

Mohamed-Amine Menacer, Abdelfetah Boumerdas, Chahnaz Zakaria, Kamel Smaili

A new language model based on possibility theory

Mohamed Amine Menacer[1], Abdelfetah Boumerdas[1], Chahnez Zakaria[1], and Kamel Smaili[2]

[1]Ecole nationale Supérieure d'Informatique, BP 68M, 16309 El Harrach, Algiers, Algeria
[2]Loria, Campus Scientifique, BP 239, 54506 Vandoeuvre Lès-Nancy, France

Abstract. Language modeling is a very important step in several NLP applications. Most of the current language models are based on probabilistic methods. In this paper, we propose a new language modeling approach based on the possibility theory. Our goal is to suggest a method for estimating the possibility of a word-sequence and to test this new approach in a machine translation system. We propose a word-sequence possibilistic measure, which can be estimated from a corpus. We proceeded in two ways: first, we checked the behavior of the new approach compared with the existing work. Second, we compared the new language model with the probabilistic one used in statistical MT systems. The results, in terms of the METEOR metric, show that the possibilistic-language model is better than the probabilistic one. However, in terms of BLEU and TER scores, the probabilistic model remains better.

Keywords: Machine Translation, probabilistic approach, the possibility theory, Language Model.

1 Introduction

Language models are essential in many domains such as Statistical Machine Translation (SMT) [2], Automatic Speech Recognition (ASR) [20], Optical Character Recognition (OCR) [23] etc. They are generally used to ensure that a system not only produces the right words but also to output them in a fluent language.

In SMT, the main role of language models is to choose, among the huge number of hypotheses, the right translations. These models are generally based on the probability theory. They proceed by estimating the probability of a linguistic event [1] from a sufficiently large text corpus.

Many language models have been proposed and developed in literature, such as: cache based models [15], trigger based models [16] and multi-gram models [6]. These three models are strongly related with the n-gram model [11], which is the most dominant and the most used model in the language modeling field. Other language models, based on other paradigms than n-gram, have been proposed. For instance, the connexionist language model, proposed by [1], which is very useful when the training data is limited.

[1] A linguistic event is the succession of one or several words.

There are various limitations of n-gram models. The classical limitations include, for example, the context-dependent issue. In fact, the probability assigned by the model is related to the context in which it occurs. This context is modeled by the size n of the model, which is, in many cases, insufficient to capture the underlying topic information. Moreover, there is another problem related to the probability theory: generally the probability assigned by these language models is significant especially with high and medium frequency events. However, when it comes to low frequency events, the theory breaks down and the model fails to assign reliable probabilities. Therefore, it turns to smoothing techniques to estimate the probability of the corresponding event.

In this paper, we present a new language modeling approach using the possibility theory [17], which is an alternative to the probability theory in order to deal with uncertainty. This paper is organized as follows: the next section briefly presents the probabilistic n-gram language models, followed by a description of the possibility theory and the motivation that encouraged us to use it. Then, the only work, in our knowledge, that has been done to model a possibilistic language model [18] is presented. Finally, we present and describe the language model based on the possibility theory along with some tests to compare its performances with those of the baseline language models.

2 Probabilistic n-gram language models

A probabilistic language model could be considered as a function that measures how likely a sentence can be expressed by a speaker in a particular language. In a more formal way, if the following word-sequences: $e = e_1 e_2 ... e_l$ is proposed by a system, then the probability $P(e)$ can be decomposed and calculated as in (1).

$$P(e) = P(e_1 e_2 ... e_l) = \prod_{j=1}^{l} P(e_j | e_1 e_2 ... e_{j-1}) \tag{1}$$

In formula (1), we notice that the probability of the word-sequence e depends on the conditional probability of each word e_j knowing its history $e_1 e_2 ... e_{j-1}$. However, the length of the taken history is a major problem because it is impossible to find all the possible histories that can precede a given word. Consequently, the exact value of the probability $P(e_j | e_1 e_2 ... e_{j-1})$ cannot be calculated. To deal with this issue, the n-grams language model assumes that the word e_j depends only on the $n-1$ words that precede it as it is shown in the equation (2).

$$P(e) = \prod_{j=1}^{l} P(e_j | e_{j-n+1} e_{j-n+2} ... e_{j-1}) \tag{2}$$

However, according to the previous equation, if the n-gram model fails to find a similar word-sequence in the training data, the probability of the n-gram could be assigned a zero value. This can be misleading because an unseen n-gram doesn't mean that it is wrong.

Generally, some other techniques are used along with the n-gram language model in order to estimate and adjust the value of the probability $P(e)$ and

to avoid the problem of assigning zero probability to unseen n-grams; such methods are called the smoothing techniques.

The smoothing techniques can be divided into two groups: the first one concerns the count smoothing methods such as the add-one smoothing, the deleted estimation and Good-Turing smoothing. The second one considers interpolation and back-off based smoothing techniques like: linear interpolation, back-off, Witten-Bell smoothing [4], Kneser-Ney smoothing [12] and Modified Kneser-Ney smoothing [4]. However, there is no technique that performs well in all situations. For all these reasons, we propose to use another theory: the possibility theory.

3 The Possibility Theory

The possibility theory was proposed by Zadeh [17] in 1978 as an extension of fuzzy sets theory and fuzzy logic. It is a mathematical framework that deals with the representation of uncertain information resulting from incomplete knowledge [9].

Traditionally, all uncertainties of information are quantified and handled by the probability theory. However, the theory of probability knows some gaps in some cases like the famous Bertrand paradox [21] where one could obtain several different values of probability for the same event only by changing the way of reasoning. The probability theory uses only one measurement to deal with uncertainty, which is the probability measure itself. Therefore, it is limited in certain situations. For example, if we affirm that the sentence *"I eat an apple"* is a well written English sentence with a probability of 0.8, we are faced with two problems. The first one is the percentage of error of this probability, which indicates the inaccuracy of respecting English language usage rules. The second issue concerns the uncertainty, i.e. there is a 80% chance that this sentence faithfully respects all the usage rules of the English language.

To take these phenomena into account, the possibility theory employs two measures: the necessity N and the possibility Π. The first allows to affirm that an event is possible with a certain degree and the second represents the degree of certainty of an event. Consequently, the possibility theory completes the probability theory.

In the same way, as in the probability theory, where the probability P can be obtained from the probability distribution, the possibility and necessity measures can be defined from the possibility distribution π in the possibility theory. But unlike the probability theory, the use of the possibility and the necessity measure in the possibility theory allows us to distinguish what is plausible from what is less plausible, what is normal from what is not, what is surprising from what is expected [8]. These advantages can be very useful to build a possibilistic language model that can reduce some limitations of the probabilistic language models.

The possibility distribution is a function of the universe Ω to the interval unit $[0, 1]$ with the following interpretations: $\pi(s) = 1$ means that the event s is possible and $\pi(s) = 0$ means that the event s is considered as impossible.

In the finite state, the possibility and the necessity measures are defined as described in (3) and (4).

$$\Pi(A) = \max_{s \in A}(\pi(s)) \tag{3}$$

$$N(A) = \min_{s \notin A}(1 - \pi(s)) \tag{4}$$

where A is a subset of the reference set Ω, representing a set of events.

In summary, the uncertainty of a proposition or an event A in the possibilistic model is measured by the couple $(\Pi(A), N(A))$, unlike the probabilistic model where the uncertainty of an event is measured by only the probability P.

4 Related works

To our knowledge, the only language model based on the possibility theory is the one described in [18]. This work uses the web as an open training corpus to improve language modeling. The reason behind using the web lies in the great and constantly growing volume of its textual documents. The resulting language model was integrated in a speech recognition system.

The basic idea behind the method of Oger and Linarès [18] is to consider the observed events and non-observed ones in the web. Therefore, two theories were employed: probability and possibility. The probability theory is used to take advantage of events observed on the Web, and the possibility theory is used to take advantage of unobserved events on the web.

To benefit from both theories: probability and possibility, Oger and Linarès [18] also proposed a method to combine the probability and the possibility measures.

5 Possibility estimated on a text corpus

The method proposed by Oger and Linarès [18] estimates the possibility of a word-sequence from the web or from a text corpus. The basic idea of this method is the following: for a given word-sequence, the possibility is estimated according to its sub-sequences. Consequently, the more sub-sequences of the original word-sequence exist in the training corpus, the higher this word-sequence is possible. A similar principle is used in the BLEU measure, which is used to evaluate the quality of a translation [19]. Thus, for a higher reliability of results, the size of the sub-sequences is given according to the order of the model. Every time a n-gram sub-sequence does not exist, the existence of all its $(n-1)$-grams sub-sequences is studied until a uni-gram level is reached.

This idea is formulated for a n-gram language model following a recursive equation. This equation determines a set of possibility distributions starting from π_n until reaching π_1:

$$\pi_n(W) = \begin{cases} \frac{|W_n \cap C_n| + \alpha |W_n / C_n|}{|W_n|} & \text{if} \quad n \geq 1 \\ 0 \qquad \text{else} \end{cases} \tag{5}$$

where:

- W: is a sequence of one or more words.
- W_n: is a set of word sequences of size n in W.
- C_n: is the set of word sequences of size n in the corpus C.

– α: is the back-off coefficient with $0 \leq \alpha \leq 1$.
– /: is the set subtraction operator.

This formula takes into account, at the same time, the observed and the unobserved events. To calculate the possibility for a word-sequence W, the number of sub-sequences of size n of W present in the training corpus is normalized by the total number of sub-sequences of size n in W. Thus, to propose a precise possibility distribution, a back-off strategy is used recursively to interpolate high order possibilities from its sub-sequence possibilistic scores.

The possibility distribution defined above allows deriving the possibility measure as in equation (6).

$$\Pi_n(A) = \max_{W \in A}(\pi_n(W)) \tag{6}$$

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

If the whole corpus is considered as a unique element W, then:

$$\Pi_n(W) = \pi_n(W) \tag{7}$$

6 A new approach based on the possibility theory

The approach proposed by Oger and Linarès [18] has some limits. For instance, when calculating the possibility of a word-sequence in their formula, the sequence is initially divided into a group of n-grams, then each n-gram is checked whether it exists in the training corpus or not. By doing so, all the extracted n-grams will have the same weight. This could be considered as a limit for the language model. In fact, even if a sequence of words belongs to a foreign language [1], the above formulation could give it a high possibility.

To overcome this limit, a novel approach, based on the same idea as that of Oger and Linarès [18] is proposed in our study. In fact, Oger and Linarès [18] define the possibility of a word-sequence based on the existence or the non-existence of this sequence or its sub-sequences of size n in the training corpus. However, in our approach, we consider a weight estimated from a training corpus, which reinforces the possibility of a word-sequence that contains a subsequences often encountered. Furthermore, we interpolate the possibility of a sequence with that of the sequences of lower size. This allows, on one hand, to take into account the case where a sequence does not appear in the training data. On the other hand, it increases the possibility distribution. By considering these assumptions, the possibility-based language model estimates the possibility of a sequence of n words as follows:

$$\pi_n(W) = \sum_{k=1}^{n} \lambda_k \alpha_k \sum_{i=1}^{\alpha} \frac{N(w_i^k)}{\beta_k} \tag{8}$$

where:

– W_i^k: is the i-th sub-sequence of size k in W.

[1] This is possible for close languages which share several words such as English and French.

- $N(W_i^k)$: is the number of occurrences of the word w_i of size k in the training corpus C.
- α: is the number of unique sub-sequences w_i of size k in the test corpus W.
- β_k: is the total number of sub-sequences of size k in the training corpus C.
- $\alpha_k = \frac{\alpha'}{\alpha}$: where α' is the number of units of size k in the test corpus that exist in the training corpus where $\alpha' \leq \alpha$.
- λ_k: is the possibility weighting coefficient.
- $\sum_i \lambda_i = 1$.

Suggesting that $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ will ensure that a more important possibility is assigned to longer sequences.

For a word-sequence W, its possibility value is defined according to the sub-sequences composing it. For each sub-sequence of size k, we determine its ponderation estimated from the training corpus. Additionally, we back-off to smaller possibilistic model. This back-off is handled by the two coefficients λ_k and α_k.

To determine the value of the possibility measure for a test corpus C_t made up of several word-sequences of size n or more, we apply the formula (9).

$$\Pi_n(W) = \max_{W \in C_t} (\pi_n(W)) \tag{9}$$

If the test corpus is considered such as a long sentence W, then:

$$\Pi_n(W) = \pi_n(W)) \tag{10}$$

As in the method of Oger and Linarès [18], the calculation of the possibility measure does not require the decomposition of the total possibility in conditional possibilities, which allows us to evaluate a word-sequence totally and not sequence by sequence as done with a probabilistic-based language models.

7 Tests and Results

In order to evaluate the performance of the possibilistic-language model, we consider two types of tests: the goal of the first test is to study the behavior of the new approach compared to that proposed by Oger and Linarès [18]. In the second one, we compare our approach with the probabilistic model used by the Moses decoder during the translation process. This probabilistic language model is calculated using the KenLM language modeling tool [10].

In the following section, the results obtained from the two tests are presented and discussed.

7.1 Experiment on language modeling

In this test, we considered the English corpus of EUROPARL [13]. This corpus is composed of the proceedings of the European Parliament since 1996. Table 1 illustrates the number of sentences and the number of words in this corpus. We subdivided the corpus in three parts as follows:

- 80% used as training data which corresponds to a vocabulary of 111 072 distinct words.

Table 1. Statistics on the EUROPARL corpus.

Language	$\|S\|$	$\|W\|$
English	2 218 201	53 974 751

– 10% used for tuning and more specifically for estimating the values of λ_k.
– The last 10% of the corpus used as test data.

We achieved several experiments concerning the possibility-language model on four other corpora. The aim of the first corpus is to study the possibility of a text that is well written in the source language (English). The two other corpora are poorly written in English and the final corpus is written in a foreign language (French). The purpose of these three last tests is to check if the possibilitic-language model estimates correctly their possibility, if so, the possibility of these two corpora should be low.

Statistics on the four test corpora are presented in the table 2.

Table 2. Statistics on the test corpora.

Corpus	$\|S\|$	$\|W\|$
Europarl	35 000	966 990
RandEuroparl	35 000	1 032 197
RandEnglish	35 011	1 031 452
Foreign	34 722	907 024

– Europarl is taken from test data (10% of the Europarl corpus).
– RandEuroparl is randomly generated from the vocabulary wordlist.
– RandEnglish is randomly generated from an English wordlist.
– Foreign is well written in French.

Before testing our possibilistic-language model, we show the influence of the interpolation coefficient λ_k on the calculation of the possibility measure. Initially, we fixed the values of λ_k in order to make the sub-sequence of size n more possible than the one of size $n-1$ and so on. Then, we estimated the values of λ_k using a greed search algorithm on the development corpus. Table 3 illustrates the total number of sentences and words, as well as the size of the vocabulary.

Table 3. Statistics on the development corpora.

Language	$\|S\|$	$\|W\|$	$\|V\|$
English	200 773	5 605 794	45 849

The values of λ_k fixed by hand and evaluated by a greed search algorithm are shown in table 4. The values of the possibility corresponding to these λ_k and for $n = 4$ are given in table 5.

Table 4. Values of λ_k fixed by hand and estimated by a greed search algorithm.

n	λ_k	λ^{Hand}	λ^{Calc}
4	λ_1	0.1	0.23
	λ_2	0.2	0.24
	λ_3	0.3	0.26
	λ_4	0.4	0.27
3	λ_1	0.2	0.32
	λ_2	0.3	0.33
	λ_3	0.5	0.35
2	λ_1	0.4	0.49
	λ_2	0.6	0.51

Table 5. The possibility-language model values with estimates and fixed λ_k.

Corpus	$\Pi_n^{Hand}(w)$	$\Pi_n^{Calc}(w)$
Europarl	0.33	0.46
RandEuroparl	0.01	0.03
RandEnglish	0.03	0.06
Foreign	0.005	0.01

We can notice that the values of the possibility measure obtained after the estimation of the values λ_k are a bit higher compared to the values obtained by fixing λ_k by hand. This is because of the high values of λ_k assigned to the n-grams of lower order (uni-grams and 2-grams), which are generally more frequent than the n-grams of higher order (3 and 4-grams).

Next, we compared the values of the possibility given by our language model and the one proposed by Oger and Linarès [18]. The results of this comparison are illustrated in table 6.

Table 6. Comparison between our language model and the one proposed by Oger and Linarès [18].

Corpus	n	$\Pi_n(W)$	$\Pi_n^{Oger}(W)$
Europarl	4	0.45	0.49
	3	0.6	0.77
	2	0.77	0.95
RandEuroparl	4	0.02	1.52×10^{-7}
	3	0.04	1.52×10^{-5}
	2	0.06	1.52×10^{-3}
RandEnglish	4	0.06	3.56×10^{-7}
	3	0.08	3.56×10^{-5}
	2	0.12	3.56×10^{-3}
Foreign	4	0.012	3.11×10^{-4}
	3	0.017	4.07×10^{-3}
	2	0.026	0.07

We notice that in both approaches the possibility value is low for the test corpora, which are badly written in English (the two corpora RandEuroparl and RandEnglish) or written in a foreign language. For the test corpus that is well written in English (Europarl Corpus), the value of the possibility is high in both approaches. We notice also that for the two approaches, the value of the possibility increases with the reduction of the model size. This is due to the n-grams of the lower order (1 and 2-grams) which are generally frequent in the training corpus.

7.2 Experiment on Machine Translation

In the second test, we achieved a comparison between our possibilistic language model and the probabilistic model within machine translation systems. In order to do that, we train a French to English translation model using Moses decoder [14] and its tools. Some statistics about the data used to train the translation model are illustrated in table 7.

Table 7. EUROPARL French-English bilingual corpus statistics.

Language	$\|S\|$	$\|W\|$	$\|V\|$
French	1 579 312	48 576 991	128 051
English	1 579 312	53 974 751	111 072

In order to compare between the two models, a corpus test of 1000 sentences ($\sim$ 33000 words) has been selected. The translation process uses the possibilistic-language model and then a baseline language model (probabilistic). To achieve this goal, we proceeded as follows:

1. We used Moses decoder and its toolkit to translate the sentences and associate to each one a list of its 1000 best translations proposed by the decoder.
2. For each sentence, we sorted its 1000 best translations, in a decreasing order, using the probabilistic model.
3. For each sentence, we kept only its best translation.
4. Finally, we used the MultEval tool [5] to calculate three scores: BLEU [19], METEOR [7] and TER [22] in order to evaluate the quality of the translations.

The previous steps have been used to evaluate the probabilistic language model. In order to evaluate the possibilistic-language model, the same steps (from 1 to 4) are used except for the second one where we sorted the sentences using our possibilistic model. The results concerning the two models are illustrated in table 8.

Table 8. Metric scores (BLEU, Meteor and Translation Error Rate -TER-) for all systems.

Metric	System Prob[1]	System Poss[2]
BLEU ↓	28,6	27,5
METEOR ↑	32,1	33,2
TER ↑	52,9	57,1

It is clear that the possibilistic-language model ensures the best values for METEOR compared to the baseline one. However, the two metrics BLEU and TER, confirm that the probabilistic language model achieves better results. This performance is encouraging since in [3], the authors show that the METEOR score correlates most with human judgment when MT is achieved from any language to English (which is the case in our system). However, the BLEU score correlates best when the translation is done from English into another language.

Thus, the advantage of METEOR is that it establishes correspondences between the reference and candidate translation on word matching, synonyms or words with the same root. However, the BLEU and TER scores are based on an explicit word matching between translation and reference.

Some sample translations, using the two language models, compared to the reference translation are given in table 9. These examples, among many others, show that the approach we propose achieves better translations, which could be considered as new references for the BLEU measure to help improve the quality of the translation even better.

Table 9. Example of a translation using the probabilistic language model and the possibilistic one.

source	trans Prob LM	trans Poss LM	reference
nous devons également analyser certains des enjeux créés par le traité d'Amsterdam.	we also need to look at some of the issues of the Treaty of Amsterdam.	we also have to analyse some of the issues, created by the Treaty of Amsterdam.	we also have to look at some of the challenges created by Amsterdam.
nous devons conserver ce droit afin de défendre un intérêt national.	we should retain the right to defend their national interests.	we must keep this law, in order to defend the right to a national interest.	we have to retain that right to defend a national interest.

[1] The system where probabilistic language model is used.
[2] The system where possibilistic-language model is used.

8 Conclusion

To conclude, in this paper, we presented a new language model based on the possibility theory while the majority of the language models used in speech recognition and machine translation are based on probabilistic approach. The interest of this new model is that it takes into account the uncertainty. We expect that this new method, which shows the feasibility of the principle, would be an alternative to classical language models. We have tested this method in a real machine translation system, and it achieved promising results. In fact, in terms of METEOR, our model is better than the baseline one while it is less efficient in terms of BLEU. Knowing that METEOR is more correlated with human judgments since this measure has been designed to overcome the weakness of BLEU and NIST. That is why we are optimistic about the interest of this approach in the near future.

References

1. Bengio, Y., Schwenk, H., Senécal, J.S., Morin, F., Gauvain, J.L.: Neural probabilistic language models. In: Innovations in Machine Learning, pp. 137–186. Springer (2006)
2. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. vol. 16, pp. 79–85. MIT Press, Cambridge, MA, USA (Jun 1990)
3. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: Further meta-evaluation of machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation. pp. 70–106. StatMT '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008)
4. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. pp. 310–318. ACL '96, Association for Computational Linguistics, Stroudsburg, PA, USA (1996)
5. Clark, J.H., Dyer, C., Lavie, A., Smith, N.A.: Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2. pp. 176–181. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
6. Deligne, S., Bimbot, F.: Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. vol. 1, pp. 169–172. IEEE (1995)
7. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 85–91. Association for Computational Linguistics (2011)
8. Dubois, D., Prade, H.: Possibility theory. Scholarpedia 2(10), 2074 (2007)
9. Dubois, D.: Possibility theory and statistical reasoning. Comput. Stat. Data Anal. 51(1), 47–69 (Nov 2006)
10. Heafield, K.: Kenlm: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 187–197. WMT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)

11. Jelinek, F.: Continuous speech recognition by statistical-methods. vol. 64, pp. 532–556. IEEE-Inst electrical electronics engineers inc 345 E 47TH ST, New York, NY 10017-2394 (1976)
12. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. vol. 1, pp. 181–184. IEEE (1995)
13. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit. pp. 79–86. AAMT, AAMT, Phuket, Thailand (2005)
14. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 177–180. ACL '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007)
15. Kuhn, R., De Mori, R.: A cache-based natural language model for speech recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 12(6), 570–583 (1990)
16. Lau, R., Rosenfeld, R., Roukos, S.: Trigger-based language models: A maximum entropy approach. In: Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing: Speech Processing - Volume II. pp. 45–48. ICASSP'93, IEEE Computer Society, Washington, DC, USA (1993)
17. Negoita, C., Zadeh, L., Zimmermann, H.: Fuzzy sets as a basis for a theory of possibility. Fuzzy sets and systems 1, 3–28 (1978)
18. Oger, S., Linarès, G.: Web-based possibilistic language models for automatic speech recognition. Computer Speech & Language 28(4), 923–939 (2014)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
20. Rosenfeld, R.: Two decades of statistical language modeling: Where do we go from here? p. 2000 (2000)
21. Saporta, G.: Probabilités, analyse des données et statistique. Editions Technip (2011)
22. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas. pp. 223–231 (2006)
23. Srihari, R., Baltus, C.: Combining statistical and syntactic methods in recognizing handwritten sentences. In: AAAI Symposium: Probabilistic Approaches to Natural Language. pp. 121–127 (1992)