Heterogeneous Statistical Language Model

Kairit Sirts
Tallinn University of Tehcnology, Department of Informatics
kairit.sirts@gmail.com

Abstract. We outline the proposal for the doctoral research of heterogeneous statistical language model. The conventional language models can be classified as homogeneous, because the usage of certain structures as the base language units (words or morphemes) in lexicon is fixed. We, on the contrary, propose not to constrain the language structures to be used in the lexicon beforehand. Our model would allow the insertion of morphemes, words as well as multi-word expressions into the lexicon. For finding the optimal lexicon, we propose two criteria: the amount of language covered with the lexicon and the amount of structure of the language preserved. Such structure-preserving model will hopefully lead to better results of certain Natural Language Processing applications, like for example Automatic Speech Recognition or Machine Translation.

Keywords: Statistical language modeling, Heterogeneous language model, Heterogeneous structure of language.

1 Introduction

In the field of Natural Language Processing (NLP) there are many tasks, which use statistical language modeling (SLM). For example Automatic Speech Recognition (ASR), Machine Translation (MT), Information Retrieval (IR) and Automatic Text Segmentation are just some examples, where the statistical language models are used. Most of these tasks could be (and historically also have been) approached also with non-statistical methods, but in the last three decades the usage of statistical models has become common in such applications [17].

The most common building blocks used in statistical language models are the words. This approach has been satisfactorily applied in the languages like English, which is a rather isolating language. In more synthetic languages, where from one word many different forms can be made, the usage of morphemes has proved to be more successful [1,3]. In both cases, the building blocks are chosen keeping in mind the properties of the specific language being modeled.

We want to propose the outline of the research of the so-called heterogeneous language model. In this model the structure of the modeling units would not be restricted. We assume that the information of the most optimal building blocks is present in the text corpus and we believe that it is possible to devise unsupervised methods, which automatically capture those structural language elements from the corpus that are most beneficial for constructing a statistical language model regardless the language. The only predefined limit would be the size of the lexicon used in the model. One would argue that the lexicon sizes used now in practical applications are not big enough to allow the capture of more complicated structures than words. Yet the success of statistical modeling is mainly due to the increase in computational power and in the future there will be more and more computational power available, which will allow the size of the lexicon to grow.

The rest of this paper is organized as follows: in section 2 we will give a brief overview of statistical language modeling and the basic mathematics involved in it. In section 3 we describe the multi-level structural division and one-level structural division of the language. In section 4 we give a short overview of the published works that might be relevant for the planned research. In section 5 the proposed research of heterogeneous statistical language model is outlined. There we also list the main problems and questions revealed so far. The section 6 gives a small example of the problem. The last section concludes the paper.

2 Statistical Language Modeling

The purpose of statistical language modeling is to capture the regularities present in a language into the statistical framework. The model consists of two parts: a lexicon and the probabilistic parameter space. The lexicon is simply a list of language elements, usually words. The set of parameters describe the regularities between the lexicon elements in a probabilistic manner.

The most common approach is to model the regularities between the words with the chain of conditional probabilities:

$$P(w_1 \dots w_i) = P(w_1) \cdot P(w_2|w_1) \cdot \dots \cdot P(w_i|w_1 \dots w_{i-1}). \quad (1)$$

The probability of the next word w_i is predicted on the history $w_1 \dots w_{i-1}$ observed. Usually not the whole history of a word will be used for prediction, because this would require an immense amount of parameters. It is a common method to make the Markov assumption that the probability of the next word is conditioned only on some fixed length history. For example, considering the history of two previous words only, the model will get the form:

$$P(w_1 \dots w_i) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2) \cdot \dots \cdot P(w_i|w_{i-2} w_{i-1}). \quad (2)$$

In general, such models are called n-gram models where n-1 determines the length of the history the next word is conditioned on. Thus the model, where the next word is conditioned on the last word only, is called bigram model and the model conditioned on the last two preceding words is called trigram model.

The conditional probabilities with history length equal to n-1 are calculated from the training corpus as follows:

$$P(w_i|w_{i-n+1} \dots w_{i-1}) = \frac{count(w_{i-n+1} w_{i-n+2} \dots w_i)}{count(w_{i-n+1} w_{i-n+2} \dots w_{i-1})} \quad (3)$$

One important aspect is how to measure the goodness of a specific model. The goodness can be tested by using the language model in a NLP application like ASR and the results are reported in the means of accuracy of the application, namely word error rate (WER) in ASR. In many occasions the appropriate NLP application is not available or it is too costly to integrate the language model into the application framework just for testing purposes [7]. In such cases a computational measure called perplexity is used. The common approach is to put a small amount of training corpus (usually 10%) aside for the purpose of perplexity evaluation and not to use this part for the model training. The perplexity is given with the equation:

$$pp = 2^{-\sum_x p(x) log_2 p(x)}. \tag{4}$$

In this equation the exponent is equal to the entropy of the data; p(x) over all x is the prior probability of the test data. As the real probability distribution of the test data is not known, p(x) is substituted by q(x) which is the observed probability distribution (posterior probability) of the training data. Thus the equation for calculating the perplexity of the test data gets the form:

$$\widetilde{pp} = 2^{-\sum_{i=1}^{N} \frac{1}{N} log_2 q(x_i)} \tag{5}$$

N is the number of elements in the test set and the reciprocal of N is substituted for p(x) for denoting the empirical distribution of the test data.

Perplexity measures the uncertainty that the model has towards the test data. The better the model fits to the test data distribution the less the model is uncertain about the data and thus the lower the perplexity. It is common practice to evaluate different models on the same test data set and endeavor toward models, which result in smaller perplexity.

Another measure used in statistical language modeling is the Out Of Vocabulary (OOV) rate which measures how many words of the test data were not present in the model's lexicon.

3 Structure of the Language

3.1 Multi-Level Structural Division

Language can be viewed as existing of structural elements of different size. We can organize them as different levels of language structure. The first level constitutes the smallest structural elements – characters. The next structural elements according to size are morphemes. From morphemes the words will be constructed. Words are organized into phrases, which in turn are used to make clauses and finally sentences. Sentences are organized into paragraphs, which in turn are parts of some bigger text.

It is obvious that the bigger the size of the language structure, the more there are different elements on that level. For example, there are only about 100 characters in the extended Latin alphabet, while the amount of different meaningful sentences is practically infinite.

In such a way we have constructed the multi-level structural division of a language. We started with the character level and ended with the level, where the whole texts could be considered as entities. We can consider each level of this structural imagination of language to be of *homogeneous* type. This means, that each level consists of entities with the same or similar structure: character level contains the whole alphabet, morpheme level consists solely of morphemes, word level contains only words etc. Each level will cover the language totally, when we are allowed to have a theoretical lexicon of infinite size.

3.2 One-level Structural Division

As a kind of opposition we now want to look at a structural division of a language, where there is only one level, which contains the entities that can be defined to be of *heterogeneous* type. This approach has been adopted by cognitive linguists in whose works such entities are called *constructions* (see for example [13]). Constructions can be morphemes, words, idioms, partially lexically filled and fully general linguistic patterns.

Almost the same idea is expressed in [19], where the notion Morpheme Equivalent Unit (MEU) is adopted. The definition of MEU is given: a word or word string, whether incomplete or including gaps for inserted variable items, that is processed like a morpheme.

The constructionists' partially lexically filled patterns conform to partly-fixed frames, which is just one possible type of MEU. The process of deciding whether some pattern is construction or MEU is slightly differently motivated by both cases. A pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist, while the finding of MEU-s follows the concept of Need Only Analysis (NOA), where nothing is broken down into smaller structural elements, unless there is a specific need for it. This means that the input is checked against the existing lexicon units and only where the variation is identified between the lexicon and input text, the further analysis follows, leading to the possible insertion of a new MEU into the lexicon.

4 Related Work

In addition to conventional word-based n-gram language models, among which trigram and bigram models are the most common, there are several more complex methods proposed. In our research we are most interested in the models, where 1) the n-grams are not restricted to the sequence of words observed alongside in the training corpus; 2) the grammatical structure of the language is captured; and 3) some linguistic patterns are revealed. Also we are interested in methods for automatic morpheme discovery, because this might give an insight of how to find the morpheme-like units, which may also extend over several words.

One big class constitute the models that learn the syntactical structure of the training corpus. In [2] first the sentence structure in the means of Parts-Of-Speech (POS) tags is estimated and then most likely words in places of POS tags are found. Models described in [4], [6], [8] and [20] create for each input sentence a parse-tree, which takes into account also head-words of the sentence parsed so far, enabling to capture the dependencies of the words, which are not alongside in the sentence.

The model in [5] makes use of dependency grammar, which is created separately for each training sentence. The grammars are presented by directed graphs, showing which words are dependent on each other.

In [12] instead of the conventional n-grams the multigrams are used. Fixed m is given and all n-s in n-grams must satisfy the inequality n <= m.

The model in [15] is trained using Latent Semantic Analysis (LSA). The words are also provided with their POS-tags, the latter information being used in the training process. In the training also the information about content words and stop words is used, helping to catch the dependencies between the words more accurately.

The methods for expressing the dependencies of the words not been alongside in the training corpus are presented in [16] and [18]. In [16] those dependencies are called triggers, while in [18] they are called intermediate distance n-grams. The essential nature of the both constructions is the same; they are basically n-grams, in which certain number of words is allowed to be between the words the n-gram consists of.

The models with lexicon consisting of morphemes instead of words are presented for example in [1], [3] and [10]. The methods for unsupervised discovery of morphemes from a text corpus are described in [9] and [14]. In [11] the morphemes discovered from a corpus are used for finding the multi-word structural patterns from the language.

5 Heterogeneous Statistical Language Model

The essence of the thesis will be established on the heterogeneity of the language structure. We will make a kind of simplification and from this point forward we will use the notation of MEU to refer to any structural language pattern, be it morpheme, word or a multi-word expression.

The planned scope of the research includes extracting the lexicon consisting of MEU-s from a text corpus and building the statistical language model. Two main questions have to be answered:

- How to find and extract the relevant MEU-s from the text corpus?
- How to build a statistical language model based on this lexicon?

We now inspect the both problems a little bit closer and we start with the problem of segmenting the text corpus into heterogeneous entities MEUs. First it must be pointed out that the objective of segmenting the corpus is to extract a certain amount of elements to be put into the lexicon of a statistical language model. The size of the lexicon can vary to some extent, but for practical reasons it must stay into certain borders, which, as already mentioned, are usually between 50K and 120K items. As we already explained, the bigger structural elements we consider as lexicon items, the more language structure will be stored into the lexicon; alas the bigger lexicon size is necessary for covering the same amount of language and preserving the low OOV rate. In a situation where the lexicon size is fixed in terms that its upper limit is given, two degrees of freedom remain – the amount of structure preserved and the amount of language covered. Now we come to some questions or sub-problems:

- How to measure the amount of preserved structure with a certain lexicon selection?
- How to decide that the balance between structure and coverage in circumstances of certain lexicon size has been achieved?

- How to achieve this balance between the preservation of structure and coverage of the language with the fixed lexicon size?

First we need to get the text corpus segmented into elements to be potentially put into lexicon. When building a lexicon where the items are allowed to have only homogeneous structure, this task is simple. It is easy to find the set of all distinct words of a corpus, calculate the frequencies of occurrences for each word and take the first N most frequent words to be put into lexicon. The task is a bit more complicated with morphemes as specialized algorithms are necessary for finding the morpheme borders in the words. If the morpheme sequence has been discovered the same process as with words can be carried out to form a lexicon. The task is however not trivial with heterogeneous elements, because there are no clear borders between the different elements in the text and also the elements can overlap each other in the sense that a word may constitute a MEU while there might be another multi-word MEU containing the same word. So we need to devise a method for finding the possible MEU-s from the text corpus. We have not yet performed any research in this matter, but the following ideas could serve as a starting point:

- The types of elements the text is to be segmented to could be restricted to morphemes, words and multiword expressions.
- The list of all such elements with their frequencies could be formed;
- There are several possible ways of how to approach to finding the multi-word expressions. For example 1) find the most frequent collocations within the sentence; 2) adopt the Part Of Speech (POS) tags to find the patterns with gaps, where some certain types of words (nouns, verbs etc) can be inserted into.

It must be noticed that the result of such an activity is not a proper segmentation of text but rather a list of textual patterns and this list has the heterogeneous structure as it may contain morphemes, words as well as multi-word expressions.

Next task is to select the proper elements from the list to be inserted into the lexicon. Due to the circumstances that the list will contain also the multi-word MEU-s, there might be several sets of items that can equally well form a lexicon and this means we need to devise a method for selecting the most optimal set of elements, which preserves the structure of the language best, while covering the most language at the same time.

Let us assume now that we have been able to extract the relevant heterogeneous entities into the lexicon which represents an optimal balance between the structure retained and language covered. As our longer goal has been to build a statistical language model then logically the question follows: how to use this lexicon for building a statistical language model? How to build up the conditional probabilities between the lexicon items considering that some of them are multi-word expressions and may even contain gaps and most probably some of them are overlapping each other?

There remains yet one important question of how to measure the goodness of this language model comparing to other statistical language models? As was explained before, there are two common methods of measuring the goodness of a language model – 1) computationally and 2) by evaluating the output of a NLP application where the model has been incorporated into. In both cases the evaluation of the model has to be set up as a comparison of the proposed model with a baseline model using some widely accepted configuration. The most common approach in statistical language modeling is to use conventional bigram and trigram models as baseline models. The computational evaluation would then involve forming the proposed model and appropriate baseline models based on the same training data and calculating the perplexity measures for all models for training and also some unseen test data.

Different NLP applications set different specific requirements to the statistical language model to be used. Therefore, in order to evaluate the goodness of a model according to the output of an application, it is necessary to choose a specific application the statistical language model has to be adapted to. In addition, as one of the targets of our research is to devise a language-independent method for finding the proper heterogeneous entities that could be used as building blocks in the statistical language model, the validation of the model should involve experiments with different languages.

6 Example

We now try to give some examples of how the heterogeneous segmentation of an English sentence could look like. As we currently do not yet have any method implementation available, which would segment the text into heterogeneous entities, the given examples present just the idea of what kind of segmentation we would expect to emerge. We also have to remind that the aim is to produce a data-driven segmentation method, which is by no means objective and is solely dependent on the data it is applied to. One of the input parameters to the system is planned to be the maximum number of elements the vocabulary can contain. Thus, the resulting segmentation will also be dependent on this parameter. For example, if the capacity of the vocabulary is for some reason only as much as 30 elements then if the corpus is Estonian text, we can put only single characters into the lexicon. While the amount of the lexicon grows, the more morphemes could be put into the lexicon, until at some point most frequent words will emerge as vocabulary element candidates. The moment when bigger structural elements will start to enter into the vocabulary is different with different languages, making the result of different languages structurally very different.

As an example we will look now the following English sentences:

Everything was going by the plan. Dinner was ready at eight as expected.

The possible segmentation of those sentences following our intuition is given below. The result is presented so that after each segment the structural type of the segment is given in brackets. The structural type phrase should not be taken literally, as we use it to describe any parts longer than words.

Everything [word] **was** [first part of the phrase] **go** [morpheme going into the phrase gap] **ing** [second part of the phrase] **by the plan** [phrase]. **Dinner** [word] **was** [word] **ready** [word] **at** [word] **eight** [word] **as expected** [phrase].

We must note that even this short segmentation is definitely arguable, because some of the parts segmented as words could be segmented as phrases with slots: was [part of the phrase] ready [word to be put into the phrase gap], at [part of the phrase] eight [word to be put into the phrase gap]. It is very difficult to estimate, how these things would work out in a real experiment. The ambiguity of the possible results speaks in the favor of the argument that the true quality of the method will be revealed only through an application.

7 Conclusion

We presented an outline for a doctoral research, which objective is to construct a statistical language model of heterogeneous language entities, by heterogeneousity meaning here that the language elements used could be any of morphemes, words or multi-words expressions. In order the reach the objective, the following problems must be tackled:

1. Find the list of heterogeneous elements of a text corpus;
2. Create an optimal lexicon of fixed size by finding the optimal balance between the preservation of the language structure and coverage of the language.
3. Use this lexicon for building a heterogeneous statistical language model.

There are several advantages we see the proposed model could have over the conventional statistical language models. First, with such model there would be no need to state explicitly beforehand of what base units the model will be constructed. The set of optimal base units to be put into the lexicon would be automatically detected based on the training corpus and structural preference of the items will probably be different for different languages. It might as well happen that with some languages the bordering situation will occur, where the optimal lexicon will consist only of items of homogeneous structure, but with the proposed model, there would be no need for a prior language-specific decision for that.

The another major possible advantage we see, is that the more structure the model lexicon will preserve while at the same time having the high coverage of the language, the better results are expected to be output by a NLP application.

References

1. Alumäe, T.: Methods for Estonian Large Vocabulary Speech Recognition. Ph. D. Thesis. Tallinn University of Technology (2006)
2. Amaya, F., Benedi, J. M.: Improvement of a Whole Sentence Maximum Entropy Language Model Using Grammatical Features. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistic, pp. 10--17 (1981)
3. Arisoy, E., Saraclar, M., Hirsimäki, T., Pylkkönen, J., Alumäe, T., Sak, H.: Mihelič, Fr., Žibert, J. (eds.) Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages. Speech Recognition : Technologies and Applications, Ch. 10, pp. 194--204 (2008)
4. Chelba, C.: A Structured Language Model. Computer Speech and Language, 14, pp. 283--332 (1998)
5. Chelba, C., Engle, D., Jelinek, F., Jimenaz, V., Khudanpur, S., Mangu, L., Printz, H., Ristad, E., Rosenfeld, R., Stolcke, A., Wu, D.: Structure and Performance of a Dependency Language Model. In: Proceedings of Eurospeech, Vol. 5, pp. 2775--2778 (1997)
6. Chelba, C., Jelinek, F.: Exploiting Syntactic Structure for Language Modeling. In Proceedings of COLING-ACL'98, Vol. 1, pp. 225--231 (1998)
7. Chen, S., Beeferman, D., Rosenfeld, R.: Evaluation Metrics for Language Models. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 275--280, (1998)
8. Collins, M.: Head-Driven Statistical Models for Natural Language Parsing. Ph. D. Thesis. Univeristy of Pennsylvania (1999)

9. Creutz, M.: Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Ph. D. Thesis. Helsinki University of Techology (2006)
10. Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, J., Varjokallio, M., Arisoy, E., Saraclar, M., Stolcke, A.: Analysis of Morph-Based Speech Recognition and the Modeling of Out-Of-Vocabulary Words Across Languages. In: Proceedings of HLT-NAACL 2007, pp. 380--387 (2007).
11. Déjean, H.: Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. In: Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, pp. 295--298 (1998)
12. Deligne, S., Bimbot, F.: Language Modeling by Variable Length Sequencies: Theoretical Formulation and Evaluation of Multigrams. In Proceedings of ICASSP, Vol. 1, pp. 169--172 (1995)
13. Goldberg, A. E.: Constructions: A New Theoretical Approach to Language. Trends in Cognitive Sciences, Vol. 7, Iss. 5, pp. 219--224 (2003)
14. Goldsmith, J., Hu, Y., Matveeva, I.: A Heuristic for Morpheme Discovery Based on String Edit Distance. Technical Report TR-2005-4. University of Chicago (2005)
15. Kanejiya, D., Kumar, A., Prasad, S.: Statistical Language Modeling with Performance Benchmarks Using Various Levels of Syntactic-Semantic Information. In Proceedings of the 20th international Conference on Computational Linguistics, Article 1161, (2004)
16. Rosenfeld, R.: A Maximum Entropy Approach to Adaptive Statistical Language Modeling. Computer Speech and Language, Vol. 10, pp. 18--228 (1996)
17. Rosenfeld, R.: Two Decades of Statistical Language Modeling: Where Do We Go from Here. In Proceedings of IEEE, Vol. 88, pp. 1270--1278, (2000)
18. Saul, L., Pereira, F.: Aggregate and Mixed-Order Markov Models for Statistical Language Processing. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pp. 8--89, (1997)
19. Wray, A.: Formulaic Language: Pushing the Boundaries. Oxford University Press (2008)
20. Wu, J., Khudanpur, S.: Combinging Nonlocal, Syntactic and N-gram Dependencies in Language Modeling. Proceedings of Eurospeech'99, pp. 2179--2182, (1999)