SEMANTIC TEXT CLUSTERS AND WORD CLASSES – THE DUALISM OF MUTUAL INFORMATION AND MAXIMUM LIKELIHOOD

Jochen Peters

Philips Research Laboratories, Weisshausstrasse 2
52066 Aachen, Germany
jochen.peters@philips.com

ABSTRACT

Dynamically modeling the word distribution in a variety of texts is a goal with various applications. For speech recognition a dynamic unigram may efficiently be used for the adaptation of longer ranging language models. For information retrieval it may be a good starting point to predict the most characteristic words in document dependent queries. This short paper presents two approaches for adaptive unigram language models and illustrates their relation in a more general information theoretic framework.

1. INTRODUCTION

A common task in speech recognition is the adaptation of rather general models to a specific application. This may involve changing acoustic conditions (e.g. speaker and channel dependency) as well as varying linguistic contexts (application domains or subdomains, speaking style etc.). A well-known problem in language modeling is a trade-off between the amount of training data required for well-trained bi- or trigrams or even longer ranging language models (LMs) and the braodening of the context with growing training corpora. A well-trained long-ranging LM for a very specific domain or a very individual speaking style is therefore normally impossible. A common task for language modeling is thus to provide algorithms to adapt well-trained but rather unspecific LMs to the most prominent characteristics of special contexts.

This paper will shortly summarize two approaches which aim at a robust quick adaptation of unigrams instead of longer ranging LMs since these require far less data for reliable estimates of local characteristics. These techniques have been published in [4, 12] and [9]. The resulting dynamic unigrams have then been used in [10] to also quickly adapt longer ranging LMs.

Both techniques show a deeper connection than can be seen from their originally independent derivation. This will become evident when an information theoretic – rather "intuitive" – point of view is taken to reinterpret the properties of the word and text clusters in the two approaches.

2. LANGUAGE MODELING APPROACHES

2.1. General procedure

A general practice in the development of successful LM techniques may be summarized as follows:

- Starting with some intuitive concept about useful structures and quantities a LM researcher tries to fix her/his ideas in a well-defined LM formula for unigram or conditional probabilities. The free parameters of such a LM have to be calculated from a training corpus.
- Almost all LMs employ in some way or another the counts of certain "events" in the training corpus (the simplest case being counts of single words, word pairs, or word triples).
- But many models also possess some free parameters which are not immediately fixed by the training corpus. Instead these have to be optimized with respect to some target function in order to get the "best" LM within the family defined by the LM structure (and the training counts).
- To best describe the typical texts of the wanted application it is most common to maximize the likelihood of some typical target corpus. This Maximum Likelihood (ML) approach may either be applied to the training corpus itself or, if enough data can be collected for the specific application, to some independent cross validation corpus.

Some common applications of this procedure are:

- Optimization of discounting parameters of so-called backing-off Mgram-LMs [6, 8]. (Here, the observed Mgram counts are discounted by the free parameter and the thus gained "probability mass" is redistributed on a shorter ranging level. This may involve a hierarchy of backing-off steps, e.g. from trigrams to bigrams to unigrams and finally to a flat distribution.)

- Linear or log-linear interpolation of several fixed probability distributions [11, 5]. Here, the interpolation weights are optimized.
- Mgram-LMs based on word classes. Here, the degrees of freedom are not just some real-valued numbers but rather the classification: All words in the vocabulary are grouped into (disjoint) classes such as to maximize the likelihood of the target corpus.

Actually, the last example is incomplete as long as we do not specify how the classes are used in the LM. The usual class-based Mgrams (see e.g. [2, 7]) use so-called bigram classes which are optimized for bigrams of the following form:

$$p(w|v) \stackrel{\text{def}}{=} p_{\text{bigram}}(C[w] \mid C[v]) \cdot p_{\text{emit}}(w \mid C[w]) \quad (1)$$

Here, the transition from the previous word v to the predicted word w is first replaced by a class based bigram p_{bigram} (which allows a better generalization from seen to unseen events than a word based bigram). The emission component p_{emit} then describes the distribution of the words within each class. It should be clear that the classification may only be optimized after fixing this LM structure (and also the dependency of p_{bigram} and p_{emit} from the event counts).

Typical bigram classes are mostly of syntactic nature, i.e. they typically separate nouns from verbs and articles etc. as well as singualar from plural, or various verb forms from each other and so on.

3. SEMANTIC CLASSIFICATION

3.1. Semantic word classes

We are now ready to introduce the first of the two approaches which aim at a robust quick estimation of local unigram distributions (see [9]). The basic idea is to have a "set of topics" with essentially disjoint topic specific vocabularies (plus some ubiquitous words). These topic specific vocabularies will also be referred to as semantic word classes S (some of which may be necessary to provide the omnipresent words). Most texts relate to several topics and consequently contain words from several classes S. (As an example from [9] you may imagine an article about 'Volkswagen' which relates to the topics 'car' and 'Germany'.) In order to model the varying proportions of different topics contributing to each text the following unigram model was proposed:

$$p_{\text{local}}(w) \stackrel{\text{def}}{=} p_{\text{local}}(S[w]) \cdot p_{\text{global}}(w \mid S[w]) \quad (2)$$

Here, the first component describes the proportion of words from the different classes, while p_{global} reflects the globally fixed distribution of words within each class (just as p_{emit} in (1)).

We should note that this LM structure is especially suited for a robust quick estimation of the local unigram distribution since the local class distribution is easily estimated from rather small texts. The major benefit of the semantic classes is the generalization from some few words to all semantically related words, e.g. from 'Volkswagen' to 'Daimler' or 'Chrysler' or to 'Germany' or 'Wolfsburg'.

As for the bigram classes descibed in Sect. 2.1 we may automatically train and optimize the classification with respect to its intended application. Following the standard procedure we may directly maximize the likelihood of the training data. The simplest target function results if both components in (2) are modeled by the maximum likelihood estimators (assuming that the classes are already known): $p^{\text{ML}}_{\text{local}}(S) = \frac{N(S,T)}{N(T)}$ and $p^{\text{ML}}_{\text{global}}(w \mid S[w]) = \frac{N(w)}{N(S[w])}$ where $N(S,T)$ gives the count of class S in the local text T, $N(T)$ and $N(S)$ are the sizes of text T and class S, respectively, and $N(w)$ is the global count of word w in the complete training corpus. This results in the following log-likelihood of the training data:

$$LL = \sum_w \sum_T N(w,T) \cdot \log \left(\frac{N(S,T)}{N(T)} \cdot \frac{N(w)}{N(S[w])} \right) \quad (3)$$

Here, $N(w,T)$ counts the words per text, and the sums run over all words from the vocabulary and all texts in the training corpus.

The clustering algorithm now has to find *that* mapping $S[w]$ of words to semantic classes that maximizes the log-likelihood given in (3).

This is the basic approach proposed in [9]. (How this is actually implemented is not our current scope.) We should just shortly mention that the nature of the thus found semantic word classes is fundamentally different from the rather syntactic bigram classes discussed before: Corresponding to our intuitive starting point most classes indeed combine words from a certain topic or domain (across all types of syntactic classes).

3.2. Semantic text clusters

The second approach is also based on a clustering technique. Now, however, the classification is not applied to words. Instead, texts from the training corpus are grouped together into "semantically coherent" subcorpora. The underlying concept is the idea that normal corpora cover a more or less broad range of different topics. (Even seemingly well-defined domains may be subdivided into subdomains, but the advantages of topic separation are greatest for rather general corpora.)

The most striking effect for different topics or domains are characteristic preferences of certain specialized vocabularies. More generally, the word distribution of each domain differs markably from the overall unigram of the combined

(mixed) corpus. An automatic subdivision of the mixture into clearly distinct subcorpora may thus help to build specialized LMs for each isolated topic. To exploit this idea for the clustering process we again start with a target function to be maximized. One starting point in [12] is again a maximum likelihood approach. Assuming that the text clusters C are already known we assign to each of them a ML unigram of the form $p^{\mathrm{ML}}_{\mathrm{cluster}}(w) = \frac{N(w,C)}{N(C)}$ where $N(w,C)$ and $N(C)$ specify the frequency of word w in C and the size of C, respectively. Evaluating each text with its own cluster unigram we arrive at the following log-likelihood of the training corpus:

$$LL = \sum_w \sum_C N(w,C) \cdot \log \frac{N(w,C)}{N(C)} \qquad (4)$$

The sums now run over all words from the vocabulary and over all text clusters.

The text clustering algorithm now has to find *that* grouping of all texts into distinct clusters that maximizes (4).

As expected from the initial concept the thus built text clusters indeed reflect semantic coherence. (Depending on the number of trained clusters some of them may be composed of more than one topic but each topic tends to be located in one or two clusters.) Within the Wall-Street-Journal corpus we could for example separate domains such as 'financial news', 'prominent people', or 'health care' etc.

Once trained the clusters are normally used as follows:

- We train one LM per cluster. To reduce smoothing problems from too small subcorpora we also train one global LM on the complete corpus.
- For new applications these LMs are then linearly interpolated with automatically optimized interpolation weights [11]. This may apply to either a static combination for a new product. (Here, the likelihood of some typical sample corpus is maximized.) Another possibility is a dynamic adaptation of the weights during a recognition process.

We should shortly mention that here – as opposed to the semantic *word* classes – the LM type is not restricted to unigrams. Each subcorpus may well serve as basis for a full Mgram model, and these Mgrams are then interpolated.

A special application of *both* approaches is a dynamic unigram adaptation. This may be interpreted as a semantic cache component which traces the current topics during a recognition process. Such a dynamic unigram may then be imposed as a dynamic marginal constraint on a longer-ranging but more general LM. In [10] a special quick adaptation scheme to such dynamic marginals was proposed which thereby allows to "distort" the general LM towards the changing semantic focus.

4. COMBINED VIEW: MUTUAL INFORMATION

Comparing the two approaches presented in Sect. 3.1 and 3.2 they first appear quite unrelated. The basic concepts and the derived formal LM structures are quite different. Furthermore, the log-likelihood expressions in (3) and (4) look completely different.

The only common concept is the exploitation of some sort of semantic coherence: The words contained in the semantic classes have a strong tendency to be concentrated in clear subsets of texts and to be absent in most other texts. The texts in some cluster on the other hand have a clear preference of some specialized subvocabulary and rarely use many words which are more specific for topics not represented in the respective cluster.

As we will see now, this semantic coherence in both approaches hints to a unifying formalism which also allows a more intuitive description of the word and text clusters' nature. Let's reformulate the just mentioned common concept:

- The starting point in both cases is a collection of texts T composed of words w. The counts $N(w,T)$ may be thought of as organized in a matrix where each line represents a word and each column stands for one text.
- After the clustering process is completed either some words w are combined into one semantic class S or several texts T are grouped together into one cluster C. This results in a count matrix $N(S,T)$ or $N(w,C)$ where either some lines or some columns of the initial matrix are combined.

Regarding (S,T) or (w,C) as two-valued random variables we will immediately see that the exploitation of semantic coherence translates into a high mutual information in the distribution given in either matrix $N(S,T)$ or $N(w,C)$.

4.1. Semantic word classes

Let's start with the semantic word classes (a fully analogous line of arguments for text clusters is given in Sect. 4.2):

1. The tendency of words from one class S to appear in some texts and to be absent from all others means that the count distribution for one fixed class S over all texts deviates significantly from a flat distribution, i.e. the counts $N(S,T)$ are typically *not at all* proportional to the text sizes $N(T)$.
2. Reversely, the non-uniform "topic composition" of each text means that the class distribution $\frac{N(S,T)}{N(T)}$ within one fixed text T is *not at all* proportional to the global class distribution $\frac{N(S)}{N}$. (Here, $N = \sum_T N(T) = \sum_S N(S)$ is the total corpus size).

In the LM model from (2) this is exploited by the adaptive component $p_{\mathrm{local}}(S)$.

These points may be expressed in several ways:

- Point 1. means that we have a high Kullback-Leibler distance between most of the normalized distributions $\frac{N(S,T)}{N(S)}$ and the relative text size distribution $\frac{N(T)}{N}$.
- Reading $\frac{N(S,T)}{N(S)}$ as $p(T \mid S)$ we may also say that the knowledge about which classes are present in some text is a good indicator for predicting the text identity.
- Combining 1. and 2., we can characterize the non-uniform class distribution over the texts as strong interdependence of the quantities S and T in the two-valued random variable (S,T).

  In an information theoretical framework, this translates into a high mutual information between S and T:

$$I = \sum_S \sum_T \frac{N(S,T)}{N} \cdot \log \frac{N(S,T) \cdot N}{N(S) \cdot N(T)} \qquad (5)$$

  For the sake of completeness we note that I may be interpreted as a weighted average over the Kullback-Leibler distances D mentioned above:

$$I = \sum_S \frac{N(S)}{N} \cdot D\left(\frac{N(S,T)}{N(S)} \,\middle\|\, \frac{N(T)}{N}\right) \qquad (6)$$

The crucial link between formalism depicted here and the LM framework from Sect. 3.1 is the dependency of both LL from (3) and of I from (5) on the wanted classification. Splitting off the terms which do *not change* with S we get:

$$\begin{aligned} LL &= \sum_S \sum_T N(S,T) \cdot \log \frac{N(S,T)}{N(S)} + \mathrm{const}_{LL} \\ N \cdot I &= \sum_S \sum_T N(S,T) \cdot \log \frac{N(S,T)}{N(S)} + \mathrm{const}_{N \cdot I} \end{aligned}$$

4.2. Semantic text clusters

The arguments for the text clusters are rather similar:

1. The tendency of the texts in one cluster C to concentrate on a specialized subvocabulary and to rarely use other topic specific words means that the word distribution within one fixed text cluster C deviates significantly from the global unigram $\frac{N(w)}{N}$.
2. Reversely, the localization of the specialized subvocabularies in the semantically related text clusters means that the distribution of one fixed word w over the text clusters C is typically *not at all* proportional to the cluster sizes $N(C)$.

As before this may be expressed in various ways:

- Point 1. means that we have a high Kullback-Leibler distance between most of the cluster unigrams $\frac{N(w,C)}{N(C)}$ and the global unigram $\frac{N(w)}{N}$.
- Reading $\frac{N(w,C)}{N(C)}$ as $p(w \mid C)$ we may say that knowing in which text clusters some word appears with high probabilities heavily narrows the word identity.
- Combining 1. and 2., we can characterize the non-uniform word distribution over the clusters as strong interdependence of the quantities w and C in the two-valued random variable (w,C).

  This again translates into a high mutual information between w and C:

$$I = \sum_w \sum_C \frac{N(w,C)}{N} \cdot \log \frac{N(w,C) \cdot N}{N(w) \cdot N(C)} \qquad (7)$$

  As for the word classes we may write I as a weighted average over the unigram Kullback-Leibler distances:

$$I = \sum_C \frac{N(C)}{N} \cdot D\left(\frac{N(w,C)}{N(C)} \,\middle\|\, \frac{N(w)}{N}\right) \qquad (8)$$

The link with the LM framework from Sect. 3.2 is now given by the cluster dependency of LL in (4) and I in (7):

$$\begin{aligned} LL &= \sum_w \sum_C N(w,C) \cdot \log \frac{N(w,C)}{N(C)} + \mathrm{const}_{LL} \\ N \cdot I &= \sum_w \sum_C N(w,C) \cdot \log \frac{N(w,C)}{N(C)} + \mathrm{const}_{N \cdot I} \end{aligned}$$

4.3. Relation between word and text clusters

Comparing the arguments and formalisms presented in the previous subsections 4.1 and 4.2 the reader will easily observe the close similarity of all steps. The only difference is an exchange between the roles of words and texts. Actually, everything may be mapped onto its counterpart by the following replacements:

General	Example
$w \leftrightarrow T$	$\frac{N(w)}{N} \leftrightarrow \frac{N(T)}{N}$
$S \leftrightarrow C$	$N(S) \leftrightarrow N(C)$
exchange	$N(S,T) \leftrightarrow N(w,C)$

This observation may be employed to use the same clustering techniques for both tasks. Starting with the matrix $N(w,T)$ or with the transposed matrix $N(T,w)$ and clustering for the first component (i.e. by lines) will either result in semantic word classes or in semantic text clusters.

4.4. Possible extension

Up to now the clustering has been employed on either word or text level. As a straightforward extension we might also combine the clustering to both dimensions aiming at a high mutual information of the two-valued variable (S, C) or maximizing the likelihood of an appropriate LM. Intuitively we should expect to thus further improve the quality of both classifications.

5. CONCLUDING REMARKS

This paper has presented a compact summary of various techniques and concepts used in language modeling. As a special application semantic clustering techniques both on word and on text level have been discusssed from various points of view. It should be mentioned that such a variety of approaches for the same final formalism may always be useful for extensions and modifications of an existing algorithm.

For the sake of completeness we should note that a third technique for the exploitation of semantic correlations inherent in the $N(w, T)$ distribution has developed over the last years: This so called Latent Semantic Analysis approach – first published in [1] – operates on both word and text level simultaneously and is essentially based on a Singular Value Decomposition of the $N(w, T)$ matrix. The most prominent semantic correlations are then derived from a projection formalism using the eigenvectors with highest eigenvalues.

For the special purpose of information retrieval the presented unigram adaptation techniques might help as follows:

- In [13] and [3] it was proposed to approach the information retrieval problem in an "inverse direction": Instead of ranking all documents by some estimated "relevance" to the given query they essentially build a LM per document and then evaluate the presented query by all LMs. Using Bayes rule the ranking is now performed by the query probabilities given the individual documents.
- A crucial point is to find a suitable method for a robust LM training from rather limited single documents.
- Here, robust methods using the generalization power inherent in the semantic correlations between related words may help to build unigram LMs from very small documents.
- This may be especially useful for short types of queries which should be treated as "bags of words" rather than as well-structured texts following normal Mgram statistics.

Acknowledgements

The author would like to sincerely thank Reinhard Kneser and Dietrich Klakow for many fruitful discussions about the current topic and some of the presented relationships.

6. REFERENCES

[1] R. R. Bellegarda et al. A Novel Word Clustering Algorithm Based on Latent Semantic Analysis. In *Proc. ICASSP*, volume I, pages 172–175, Atlanta, GA., May 1996.

[2] P. F. Brown at al. Class-based n-gram models of natural language. In *Comp.Ling.*, volume 18, pages 467–479, 1992.

[3] A. Berger and John Lafferty. Information Retrieval as Statistical Translation. In *Proc. ACM SIGIR*, pages 222–229, Berkeley, CA., 1999.

[4] D. Carter. Improving Language Models by Clustering Training Sentences. In *Proc. 4th Conf. on Appl. Natural Lang. Proc.*, pages 59–64, 1994.

[5] D. Klakow. Log-linear interpolation of language models. In *Proc. ICSLP*, volume 5, pages 1695–1699, Sydney, Australia, December 1998.

[6] R. Kneser, U. Essen, and H. Ney. On the use of the leaving-one-out method in statistical language modelling. In *Proc. of NATO ASI: New Advances and Trends in Speech Recognition and Coding*, pages 183–186, Bubiòn, Spain, June-July 1993.

[7] R. Kneser and H. Ney. Improved clustering techniques for class-based statistical language modelling. In *Proc. EUROSPEECH*, pages 973–976, Berlin, Germany, Sep. 1993.

[8] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, volume 1, pages 181–184, Detroit, MI, May 1995.

[9] R. Kneser and J. Peters. Semantic clustering for adaptive language modeling. *Proc. ICASSP*, 2:779–782, Apr. 1997.

[10] R. Kneser, J. Peters, and D. Klakow. Language model adaptation using dynamic marginals. *Proc. EUROSPEECH*, 4:1971–1974, Sep. 1997.

[11] R. Kneser and V. Steinbiss. On the dynamic adaptation of stochastic language models. In *Proc. ICASSP*, volume II, pages 586–589, Minneapolis, MN, Apr. 1993.

[12] J. Peters. Document clustering for the improvement of language models. In *Proc. ITG-Fachtagung für Sprachkommunikation*, pages 59–63, Frankfurt, Germany, Sep. 1996.

[13] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proc. ACM SIGIR*, pages 275–281, Melbourne, Australia, Aug. 1998.