A Corpus Balancing Method for Language Model Construction

Luis Villaseñor-Pineda, Manuel Montes-Y-Gómez, Manuel Pérez-Coutiño, Dominique Vaufreydaz

A Corpus Balancing Method for Language Model Construction

L. Villaseñor-Pineda[1], **M. Montes-y-Gómez**[1], **M. Pérez-Coutiño**[1], and **D. Vaufreydaz**[2]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico.
{villasen,mmontesg,mapco}@inaoep.mx

[2] Laboratoire CLIPS-IMAG, Université Joseph Fourier, France.
Dominique.Vaufreydaz@imag.fr

Abstract. The language model is an important component of any speech recognition system. In this paper, we present a lexical enrichment methodology of corpora focused on the construction of statistical language models. This methodology considers, on one hand, the identification of the set of poor represented words of a given training corpus, and on the other hand, the enrichment of the given corpus by the repetitive inclusion of selected text fragments containing these words. The first part of the paper describes the formal details about this methodology; the second part presents some experiments and results that validate our method.

Keywords: language model, lexical analysis, corpora, and lexical enrichment.

1 Introduction

The language model (LM) is an important component of any automatic speech recognition system. Its purpose is to reduce the search space in order to accelerate the recognition process. There are two kinds of language models: grammar based and statistical. The statistical LMs have the capability to use the statistical properties of language in context of two or more words. Because of this, statistical LMs are more flexible than the grammar based ones, and allow capturing situations closer to spoken language (where rules for written language are not always respected).

Statistical LMs are calculated from training corpora delimited by their vocabulary size, the treatment of unknown words, and others [3]. The size of the training corpus is an essential factor of a LM. Generally, a large corpus tends to have more contexts for each word, and thus tends to produce more accurate and robust LMs.

The construction of a corpus is not an easy task mainly because the written texts do not represent adequately many phenomenon of spontaneous speech. One way to diminish this problem is using web documents as data sources. Because many people around the world contribute to create the web documents, most of them has informal contents, and include many everyday as well as non-grammatical expressions used in spoken language. This situation allows not only the construction of very large corpora but also the creation of corpora combining good written grammatical text and free text closer to the spoken language [2, 7].

Once a training corpus is constructed from the web several questions emerge. For instance, is the obtained corpus rich enough for the specified task? are the domain

words well represented? can the corpus be enriched? In this paper, we present a methodology to respond to these questions. Basically, this methodology consists of two steps: i) a lexical analysis of the training corpus in order to identify its weaknesses relating to a given reference corpus[1], and ii) a lexical enrichment process of the training corpus focused on reducing the identified weaknesses, and obtaining a better LM.

The rest of the paper is organized as follows. Section 2 introduces formal concepts of the lexical analysis of a training corpus, and explains the identification of its bad represented words. Section 3 describes its enrichment process. Section 4 shows some experiment results that illustrate and validate our method. Finally, section 5 presents our conclusions and discusses future work.

2 Lexical Analysis of the Training Corpus

It is clear that the terms and expressions used in real dialogs considerably differ from those occurring in texts. For instance, we can expect that the frequency of occurrence of pronouns and verbs in the first and second person is not similar between a dialog among people and a written text. Therefore the aim of this analysis is to find those words having very different frequencies in two corpora (i.e. between a training corpus and a reference corpus). The identified words can be over or sub represented in the training corpus related to the reference one.

The method of lexical analysis of two corpora consist of two major stages:

1. Constructing the word probability distribution for each corpus (preprocessing stage).
2. Measuring the difference between the probability distributions of the corpora, and identifying the critical words (comparison stage)

These processes are described in the next two subsections.

2.1 Preprocessing Stage

This stage considers the creation of an index of the corpora. This index indicates the words used in the corpora, and their corresponding frequencies of occurrence in each corpus. We represent this index by an inverted file, and instrument it by a set of hash tables [4].

Once the index is built, a frequency $f_t^{C_i}$ is assigned to each word t. This frequency indicates the number of occurrences of the word t in the corpus C_i. Then, using these frequencies of the words, a probability distribution $D_i = \left\{ p_t^{C_i} \right\}$ of the words in the corpus C_i is constructed, where $p_t^{C_i} = f_t^{C_i} \Big/ \sum_{j=1}^{n} f_j^{C_i}$ expresses the probability of occurrence of the word t in the corpus C_i. Here, n is the number of words considered by the index.

[1] A *reference corpus* is a set of samples for a given interaction including the linguistics phenomenon of the domain. These corpora are obtained from real (or almost real) conditions. In our case, we built the reference corpus using the technique of the Wizard of Oz (see section 4.1).

2.2 Comparison Stage

This stage aims, at a first step, determining the general difference between the corpora. Then based on this information, identifying the specific words mainly causing this difference (i.e. the set of disparate words of the training corpus).

2.2.1 Comparison of the probability distributions

In order to measure the lexical difference between the corpora, we compare their word probability distributions $D_i = \left\{ p_t^{C_i} \right\}$. Because we are interested in the general difference regardless of the direction, we propose a comparison measure *Diff* for two distributions: the quotient of the difference area and the maximal area. This measure reflects an overall difference of the corpora and does not measures individual proportions of difference of each individual word. More detail on this measure can be found in [5].

$$Diff = \frac{A_d}{A_m} \qquad \text{difference coefficient:}$$

$$A_d = \sum_{t=1}^{n} d_t \qquad \text{difference area}$$

$$A_m = \sum_{t=1}^{n} max\left(p_t^{C_e}, p_t^{C_r} \right) \qquad \text{maximal area}$$

$$d_t = \left| p_t^{C_e} - p_t^{C_r} \right| \qquad \text{word difference}$$

If the difference coefficient between the two probability distributions tends to 1, then there exists a considerable lexical difference between the corpora. On the contrary, if the difference coefficient tends to 0, then we can conclude that the corpora are lexically similar.

2.2.2 Identification of the disparate words

A global difference between the corpora is caused essentially by the abrupt differences d_t of some individual words. We call these words *disparate*, and defined them as those with a difference noticeably greater than the typical difference. Let d_m be a "typical" value of d_t and d_s be a measure of the "width" of the distribution (see below). Then a word *t* for which $d_t > d_m + (a \times d_s)$ is identified as a disparate word. The tuning constant a determines the criterion used to identify an individual difference as noticeable.

$$d_\mu = \frac{1}{n} \sum_{t=1}^{n} d_t \qquad \text{average difference}$$

$$d_s = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (d_t - d_\mu)^2} \qquad \text{standar deviation of difference}$$

3 Lexical Enrichment of the training corpus

On the basis of the lexical analysis of the corpora (i.e. the comparison of the training and reference corpora), it is possible to determine, first, the appropriateness of the training corpus related to the reference one, and then, the set of poor represent words requiring to be enriched.

The appropriateness of the training corpus is determined by the difference coefficient. If this coefficient is closer to 0, then the word distributions of both corpora are similar, and thus the training corpus is adequate for the task at hand in accordance with the reference corpus. On the contrary, if the difference coefficient is closer to 1, then the word distributions of the corpora are very different, and there are not sufficient elements to generate a satisfactory LM.

For the situation where the difference coefficient is closer to 0, it is necessary to enrich the training corpus. The lexical analysis allows determining the set of bad-represented words (i.e. the set of disparate words). From them, the subset of sub-represented words is of particular interest to be enriched. We call them *critical words.*

Two different data sources can be used to obtain samples of the critical words, and thus enriching the training corpus. On one hand is a new group of documents obtained from the web; on the other hand is the reference corpus. Since we are interested in the creation of LMs for spoken language, and the spoken phenomenons are poorly represented in the web documents (for instance deictic and courtesy expressions; see section 4.2), we decided to use the reference corpus as data source.

Basically, our method proposes to enlarge the training corpus aggregating to it several times a set of selected phrases from the reference corpus. The following section describes the selection of these phrases and their incorporation to the training corpus.

3.1 The process of enrichment

Given the set of critical words W_c (i.e. the set of sub-represented words in the training corpus) the process of lexical enrichment of the training corpus consists of the following steps:

1. Construct the selected corpus C_s from the reference corpus C_r. This new corpus contains only those phrases from the reference corpus having one or more critical words (i.e. $C_s \subseteq C_r$ and $|C_s| \leq |C_r|$).[2] Some properties about the frequency of occurrence of its words are:

$$\forall t \in W_c : f_t^{C_s} = f_t^{C_r}$$
$$\forall t \notin W_c : f_t^{C_s} \leq f_t^{C_r}$$

2. Calculate the deficit of occurrence of each single critical word. This deficit indicates the number of times the word $t \in W_c$ must be incorporated to the training corpus C_e in order to reach its probability of occurrence in the reference corpus.

$$deficit_t = (P_t^{C_r} - P_t^{C_e}) \times |C_e|$$

[2] The notation $|C_i|$ stands for the number of phrases in the corpus C_i.

3. Determine the number of times (repetitions) the selected corpus must be aggregated to the training corpus. This number of repetitions $\hat{r}$ is calculated in order to fulfill the occurrence deficit of all critical words.

$$\hat{r} = max(R), \text{ where :}$$
$$R = \{r_t | t \in W_c\}$$
$$r_t = \frac{deficit_t}{f_t^{C_s}}$$

4. Construct the enriched training corpus C_{e+}. This step consists on aggregating $\hat{r}$ times the selected corpus to the training one. The resulting enriched training corpus satisfied the following condition: $|C_{e+}| = |C_e| + (\hat{r} \times |C_s|)$.

4 Experimental results

This section shows some experiments that validate our method. This experiments use the corpus DIME as reference corpus, and the corpus WebDIME as training corpus. The following subsections describe both corpora, and presents the results for their lexical comparison, and for the lexical enrichment of the WebDIME corpus.

4.1 Corpora description

4.1.1 The DIME corpus

The DIME corpus is a multimodal corpus that provides empiric information for studying the use and interaction between spoken language, deictic gestures and the graphical context during human-computer interaction [8]. This corpus consists of a set of dialogs corresponding to the domain of kitchen design. This domain was selected because it is simple (most people can undertake it without previous experience), has a constrained language, and allows the use of deictic gestures.

For the construction of the DIME corpus, we used a so-called *Wizard of Oz* experiment. This experiment consists of a person (the wizard) playing the role of the system, and other person (the subject) solving tasks in the domain of interest with the help of the wizard [1].

The construction, and the corresponding transcription, of the DIME corpus was performed within the context of the DIME project "Intelligent Multimodal Dialogs in Spanish" [6]. Table 1 resumes the main characteristics of this corpus.

4.1.2 The WebDIME corpus

The creation of the DIME corpus was motivated by two different purposes: on one hand, the study of multimodal human-computer interactions, on the other hand, the construction of an automatic speech recognition system. Despite their richness for the first purpose, the DIME corpus is very small to be used for obtaining a statistical LM (i.e. to be used as training corpus). This situation motives us to collect a larger corpus from the web: the WebDIME corpus.

	DIME corpus	*WebDIME corpus*
Instances of lexical forms	27459	27,224,579
Lexical forms	1110	1110
Lines	5779	4,520,513

Table 1. Main data of the DIME and WebDIME corpora

The WEbDIME corpus is a large set of phrases containing just the vocabulary for the domain of kitchen design (i.e. the same vocabulary of the DIME corpus). It was constructed from almost 30 gigabytes of Spanish web documents gathered by the CLIPS-Index web robot [7]. Basically, it consists of all the minimal blocks containing the words of the domain vocabulary found in the collected documents. The table 1 resumes the main characteristics of this corpus.

4.2 Results from the lexical comparison between DIME and WebDIME

The following bullets resume the results from the comparison of the corpora:

- The difference coefficient is equal to 0.71. It indicates an important disparity among the proportions of occurrence of the vocabulary words in both corpora. This situation predicts the construction of an inadequate LM from the WebDIME corpus for the tasks of kitchen design.
- The set of critical words represents the 2.6% of the application vocabulary (see the table 2). This words are of three main kinds:
  - *Domain words* such as “refrigerator”, “cupboard” and “stove”. This is a serious problem since these words are very common in our application.
  - *Deictic words*, for instance, “there” and “here”. This omission occurs because these words are common in a multimodal interaction but not in written texts.
  - *Courtesy expressions* including auxiliary verbs such as “can” and “would”. These expressions are regular in Spanish spoken language but are almost null in written texts.

It is important to point out that in spite of the small number of critical words (just 29 words from a vocabulary of 1110), the damage caused to the LM may be substantial because it considers all usage contexts of these words. This supposition was confirmed by the experiments (see the section 4.3).

Critical words		
ahí (*there*)	esta (*this, this one*)	ponga (*put*)
ahora (*now*)	está (*is*)	puedes (*can*)
alacena (*cupboard*)	éste (*this one*)	quieres (*would*)
alacenas (*cupboards*)	estufa (*stove*)	quiero (*would*)
aquí (*here*)	fregadero (*kitchen sink*)	refrigerador (*refrigerator*)
así (*so*)	hacia (*for*)	sí (*yes*)
bien (*well*)	mueble (*stuff*)	tenemos (*have*)
bueno (*good*)	okey (*okay*)	vamos (*lets go*)
dónde (*where?*)	pared (*wall*)	ver (*to see*)
esquina (*corner*)	poner (*to put*)	

Table 2. The set of critical words of the WebDIME corpus

Training corpus	*Perplexity*	*Bigram hit factor*	*Learned bigram*
WebDIME	203.02	2797	163624
WebDIME+	16.42	3068	164462

Table 3. Evaluation of the obtained LMs

4.3 Results from the lexical enrichment of the WebDIME corpus*

The enrichment of the corpus WebDIME was done in two steps (see the section 3.1). First, we obtained a selected corpus C_s of 3278 phrases from the DIME corpus (C_r). Then, we aggregate 402 times these phrases to WebDIME (C_e) in order to build the WebDIME+ corpus (C_{e+}).

In order to estimate the adequacy of the enriched corpus, we evaluated the coverage of the resultant LM for the given task. Basically, we consider the following well-known measures: the perplexity, the n-gram hit factor, and the number of learned bigrams [3]. The table 3 compares the LMs constructed from the WebDIME and WebDIME+ corpora. These results demonstrate that the LM obtained from the new enriched corpus is better: the perplexity decreased, and the 2-gram hit factor and the number of learned bigrams increased.

Additionally, we performed another two experiments for validating our methodology. These experiments considered different ways of enriching the training corpus.

The first experiment consisted on varying the number of repetitions the selected corpus was aggregated to the WebDIME corpus. Table 4 shows the results of this experiment. In this table, WebDIME1 is a corpus conformed by WebDIME and only one repetition of the selected phrases; WebDIME262 contains 262 repetitions of the selected corpus[3], and WebDIME800 contains 800 repetitions.

Training corpus	*Perplexity*	*Bigram hit factor*	*Learned bigram*	*Diff*
WebDIME1	60.21	3068	164462	0.72
WebDIME262	17.59	3068	164462	0.66
WebDIME+	16.42	3068	164462	0.64
WebDIME800	15.04	3068	164462	0.59

Table 4. Experiments aggregating different times the selected corpus

The results show that perplexity decreased considerably between WebDIME1 and the WebDIME+, and just a few between WebDIME+ and WebDIME800. Therefore, from table 4 it is clear that the WebDIME+ corpus maintains the best relation between cost and benefit. Additionally, the table 4 shows a strong correlation between the perplexity and the difference coefficient.

In the second experiment the selected corpus was substituted by the complete DIME corpus (i.e. the construction of the selected corpus was eliminated from the

* All LMs used in the experiments were constructed by the same technique. Also, we reserved a subset of the DIME corpus for evaluation purposes. This subset was excluded for the construction of the selected corpus.

[3] 262 is the average of the repetitions of all critical words. The proposed calculus considers the maximum instead of the average (see section 3.1).

Training corpus	*Perplexity*	*Bigram hit factor*	*Learned bigram*
WebDIMED1	121.76	2947	165124
WebDIMED402	17.59	2947	165124

Table 5. Experiments aggregating the reference corpus

procedure of section 3.1). Table 5 shows the results of this experiment. In this table, WebDIMED1 is the corpus conformed by WebDIME and one repetition of the reference corpus; and WebDIMED402 consist of WebDIME and 402 repetitions of the corpus DIME (i.e. the reference corpus).

The comparison of the results of tables 4 and 5 allows concluding that using a selected corpus is an advantageous strategy for compensating the deficit of the critical words (at least a better strategy than just aggregating the reference corpus). For instance, the results shows that perplexity was lesser and 2-gram hit factor was greater when using the selected corpus.

5 Conclusions and future work

In this paper we presented a methodology for lexical corpora enrichment focused on the creation of statistical language models. This methodology consists of two major steps: first, a lexical comparison between the training and reference corpora that allows identifying the set of critical words (sub represented words) of the training corpus; second, the lexical enrichment of the training corpus.

The proposed methodology was experimented with the DIME and WebDIME corpora. The result of this experiment was the enriched corpus WebDIME+. We demonstrated that the adequacy of this new corpus for the task at hand was better than that for the original training corpus.

Additionally, we propose a new measure, the difference coefficient, to quantify the difference between two corpora. Our experiments demonstrate that, similar to traditional measures such as perplexity, this coefficient may be used to evaluate the adequacy of a corpus to a given domain.

As future work we plan to: 1) continue the evaluation of the obtained LMs over a speech recognition system, 2) propose a iterative method for corpora enrichment based on the dynamic calculus of the critical words and pertinent stop conditions, 3) extend the corpora comparison in order to consider syntactic information (such as part of speech tags).

Acknowledgements. This work was done under the partial support of CONACYT (project 31128-A), the "Laboratorio Franco-Mexicano de Informática (LAFMI)", and the Human Language Technologies Laboratory of INAOE.

References

1. Bernsen, N., H. Dybkjaer and L. Dybkjaer. *Designing Interactive Speech Systems. From First Ideas to User Testing*. Springer-Verlag. 1998.
2. Gelbukh, A., G. Sidorov and L. Chanona. Compilation of a Spanish Representative Corpus. *International Conference on Computational Linguistics and Intelli-*

gent Text Processing CICLing02, Lecture Notes in Computer Science 2276, Springer. 2002.
3. Jurafsky, D. and J. Martin. *Speech and Language Processing*. Prentice Hall. 2000.
4. Kowalski, G. *Information Retrieval Systems: Theory and implementation*. Kluwer Academic Publishers, 1997.
5. Montes y Góméz, M., A. Gelbukh and A. López-López. Mining the News: Trends, Associations and Deviations. *Computación y Sistemas*, Vol. 5, No. 1, IPN 2001.
6. Pineda, L. A., A. Massé, I. Meza, M. Salas, E. Schwarz, E. Uraga and L. Villaseñor. The DIME Project. *Mexican International Conference on Artificial Intelligence MICAI-2002*, Lecture Notes in Artificial Intelligence 2313, Springer-Verlag, 2002.
7. Vaufreydaz, D., M. Akbar and J. Rouillard. Internet Documents : A Rich Source for Spoken Language Modeling. *Automatic Speech Recognition and Understanding (ASRU`99)*, Keystone, Colorado, USA, 1999.
8. Villaseñor, L., A. Massé and L.A. Pineda. The DIME corpus. *3er Encuentro Internacional de Ciencias de la Computación ENC-01*, Aguascalientes, México, 2001.