Web-based and combined language models: a case study on noun compound identification

Carlos Ramisch[†*] **Aline Villavicencio**[*] **Christian Boitet**[†]
[†] GETALP – Laboratory of Informatics of Grenoble, University of Grenoble
[*] Institute of Informatics, Federal University of Rio Grande do Sul
{ceramisch,avillavicencio}@inf.ufrgs.br Christian.Boitet@imag.fr

Abstract

This paper looks at the *web* as a corpus and at the effects of using *web* counts to model language, particularly when we consider them as a domain-specific versus a general-purpose resource. We first compare three vocabularies that were ranked according to frequencies drawn from general-purpose, specialised and *web* corpora. Then, we look at methods to combine heterogeneous corpora and evaluate the individual and combined counts in the automatic extraction of noun compounds from English general-purpose and specialised texts. Better *n*-gram counts can help improve the performance of empirical NLP systems that rely on *n*-gram language models.

1 Introduction

Corpora have been extensively employed in several NLP tasks as the basis for automatically learning models for language analysis and generation. In theory, *data-driven* (*empirical* or *statistical*) approaches are well suited to take intrinsic characteristics of human language into account. In practice, external factors also determine to what extent they will be popular and/or effective for a given task, so that they have shown different performances according to the availability of corpora, to the linguistic complexity of the task, etc.

An essential component of most empirical systems is the *language model* (LM) and, in particular, *n-gram language models*. It is the LM that tells the system how likely a word or *n*-gram is in that language, based on the counts obtained from corpora. However, corpora represent a sample of a language and will be sparse, i.e. certain words or expressions will not occur. One alternative to minimise the negative effects of data sparseness and account for the probability of out-of-vocabulary words is to use discounting techniques, where a constant probability mass is discounted from each *n*-gram and assigned to unseen *n*-grams. Another strategy is to estimate the probability of an unseen *n*-gram by backing off to the probability of the smaller *n*-grams that compose it.

In recent years, there has also been some effort in using the *web* to overcome data sparseness, given that the *web* is several orders of magnitude larger than any available corpus. However, it is not straightforward to decide whether (a) it is better to use the *web* than a standard corpus for a given task or not, and (b) whether corpus and *web* counts should be combined and how this should be done (e.g. using interpolation or back-off techniques). As a consequence there is a strong need for better understanding of the impacts of *web* frequencies in NLP systems and tasks.

More reliable ways of combining word counts could improve the quality of empirical NLP systems. Thus, in this paper we discuss *web*-based word frequency distributions (§ 2) and investigate to what extent "*web*-as-a-corpus" approaches can be employed in NLP tasks compared to standard corpora (§ 3). Then, we present the results of two experiments. First, we compare word counts drawn from general-purpose corpora, from specialised corpora and from the *web* (§ 4). Second, we propose several methods to combine data from heterogeneous corpora (§ 5), and evaluate their effectiveness in the context of a specific multiword

expression task: automatic noun compound identification. We close this paper with some conclusions and future work (§ 6).

2 The *web* as a corpus

Conventional and, in particular, domain-specific corpora, are valuable resources which provide a closed-world environment where precise *n*-gram counts can be obtained. As they tend to be smaller than general purpose corpora, data sparseness can considerably hinder the results of statistical methods. For instance, in the biomedical Genia corpus (Ohta et al., 2002), 45% of the words occur only once (so-called *hapax legomena*), and this is a very poor basis for a statistical method to decide whether this is a significant event or just random noise.

One possible solution is to see the *web* as a very large corpus containing pages written in several languages and being representative of a large fraction of human knowledge. However, there are some differences between using regular corpora and the *web* as a corpus, as discussed by Kilgarriff (2003). One assumption, in particular, is that page counts can approximate word counts, so that the total number of pages is used as an estimator of the *n*-gram count, regardless of how many occurrences of the *n*-gram they contain.

This simple underlying assumption has been employed for several tasks. For example, Grefenstette (1999), in the context of example-based machine translation, uses *web* counts to decide which of a set of possible translations is the most natural one for a given sequence of words (e.g. *groupe de travail* as *work group* vs *labour collective*). Likewise, Keller and Lapata (2003) use the *web* to estimate the frequencies of unseen nominal bigrams, while Nicholson and Baldwin (2006) look at the interpretation of noun compounds based on the individual counts of the nouns and on the global count of the compound estimated from the *web* as a large corpus.

Villavicencio et al. (2007) show that the *web* and the British National Corpus (BNC) could be used interchangeably to identify general-purpose and type-independent multiword expressions. Lapata and Keller (2005) perform a careful and systematic evaluation of the *web* as a corpus in other general-purpose tasks both for analysis and generation, comparing it with a standard corpus (the BNC) and using two different techniques to combine them: linear interpolation and back-off. Their results show that, while *web* counts are not as effective for some tasks as standard counts, the combined counts can generate results, for most tasks, that are as good as the results produced by the best individual corpus between the BNC and the *web*. Nakov (2007) further investigates these tasks and finds that, for many of them, effective attribute selection can produce results that are at least comparable to those from the BNC using counts obtained from the *web*.

On the one hand, the *web* can minimise the problem of sparse data, helping distinguish rare from invalid cases. Moreover, a search engine allows access to ever increasing quantities of data, even for rare constructions and words, which counts are usually equated to the number of pages in which they occur. On the other hand, *n*-grams in the highest frequency ranges, such as the words *the*, *up* and *down*, are often assigned the estimated size of the *web*, uniformly. While this still gives an idea of their massive occurrence, it does not provide a finer grained distinction among them (e.g. in the BNC, *the*, *down* and *up* occur 6,187,267, 84,446 and 195,426 times, respectively, while in Yahoo! they all occur in 2,147,483,647 pages).

3 Standard vs *web* corpora

When we compare *n*-gram counts estimated from the *web* with counts taken from a well-formed standard corpus, we notice that *web* counts are "estimated" or "approximated" as page counts, whereas standard corpus counts are the exact number of occurrences of the *n*-gram. In this way, *web* counts are dependent on the particular search engine's algorithms and representations, and these may perform approximations to handle the large size of their indexing structures and procedures, such as ignoring punctuation and using stopword lists (Kilgarriff, 2007). This assumption, as well as the following discussion, are not valid for for controlled data sets derived from Web data, such

as the Google 1 trillion n-grams[1]. Thus, our results cannot be compared to those using this kind of data (Bergsma et al., 2009).

In data-driven techniques, some statistical measures are based on contingency tables, and the counts for each of the table cells can be straightforwardly computed from a standard corpus. However, this is not the case for the *web*, where the occurrences of an n-gram are not precisely calculated in relation to the occurrences of the $(n-1)$-grams composing it. For instance, the n-gram *the man* may appear in 200,000 pages, while the words *the* and *man* appear in respectively 1,000,000 and 200,000 pages, implying that the word *man* occurs with no other word than *the*[2].

In addition, the distribution of words in a standard corpus follows the well known Zipfian distribution (Baayen, 2001) while, in the *web*, it is very difficult to distinguish frequent words or n-grams as they are often estimated as the size of the *web*. For instance, the Yahoo! frequencies plotted in figure 1(a) are flattened in the upper part, giving the same page counts for more than 700 of the most frequent words. Another issue is the size of the corpus, which is an important information, often needed to compute frequencies from counts or to estimate probabilities in n-gram models. Unlike the size of a standard corpus, which is easily obtained, it is very difficult to estimate how many pages exist on the *web*, especially as this number is always increasing.

But perhaps the biggest advantage of the *web* is its availability, even for resource-poor languages and domains. It is a free, expanding and easily accessible resource that is representative of language use, in the sense that it contains a great variability of writing styles, text genres, language levels and knowledge domains.

4 Analysing n-gram frequencies

In this section, we describe an experiment to compare the probability distribution of the vocabulary of two corpora, Europarl (Koehn, 2005) and Genia (Ohta et al., 2002), that represent a sample of general-purpose and specialised English. In addition to both corpora, we also considered the counts from the *web* as a corpus, using Google and Yahoo! APIs, and these four corpora act as *n-gram count sources*. To do that, we preprocessed the data (§ 4.1), extracted the vocabularies from each corpus and calculated their counts in our four n-gram count sources (§ 4.2), analysing their rank plots to compare how each of these sources models general-purpose and specialised language (§ 4.3). The experiments described in this section were implemented in the `mwetoolkit` and are available at `http://sf.net/projects/mwetoolkit/`.

	V_{ep}	V_{genia}	V_{inter}
types	104,144	20,876	6,798
hapax	41,377	9,410	–
tokens	39,595,352	486,823	–

Table 1: Some characteristics of general vs domain-specific corpora.

4.1 Preprocessing

The Europarl corpus v3.0 (*ep*) contains transcriptions of the speeches held at the European Parliament, with more than 1.4M sentences and 39,595,352 words. The Genia corpus (*genia*) contains abstracts of scientific articles in biomedicine, with around 1.8K sentences and 486,823 words. These standard corpora were preprocessed in the following way:

1. conversion to XML, lemmatisation and POS tagging[3];
2. case homogenisation, based on the following criteria:
   - all-uppercase and mixed case words were normalised to their predominant form, if it accounts for at least 80% of the occurrences;
   - uppercase words at the beginning of sentences were lowercased;
   - other words were not modified.

[1] This dataset is released through LDC and is not freely available. Therefore, we do not consider it in our evaluation.

[2] In practice, this procedure can lead to negative counts.

[3] Genia contains manual POS tag annotation. Europarl was tagged using the TreeTagger (`www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`).

This lowercasing algorithm helps to deal with the massive use of abbreviations, acronyms, named entities, and formulae found in specialised corpora, such as those containing biomedical (and other specialised) scientific articles.

For calculating arbitrary-sized n-grams in large textual corpora efficiently, we implemented a structure based on suffix arrays (Yamamoto and Church, 2001). While suffix trees are often used in LM tools, where n-grams have a fixed size, they are not fit for arbitrary length n-gram searches and can consume quite large amounts of memory to store all the node pointers. Suffix arrays, on the other hand, allow for arbitrary length n-grams to be counted in a time that is proportional to $\log(N)$, where N is the number of words (which is equivalent to the number of suffixes) in the corpus. Suffix arrays use a constant amount of memory proportional to N. In our implementation, where every word and every word position in the corpus are encoded as a 4-byte integer, it corresponds precisely to $4 \times 2 \times N$ plus the size of the vocabulary, which is generally very small if compared to N, given a typical token/type ratio. The construction of the suffix array takes $O(N \log_2 N)$ operations, due to a sorting step at the end of the process.

4.2 Vocabulary creation

After preprocessing, we extracted all the unigram surface forms (i.e. all words) from *ep* and from *genia*, generating two vocabularies, V_{ep} and V_{genia}, where the words are ranked in descending frequency order with respect to the corpus itself seen as a n-gram count source. Formally, we can model a *vocabulary* as a set V of words $v_i \in V$ taken from a corpus. A word *count* is the value $c(v_i) = n$ of a function that goes from words to natural numbers, $c : V \to \mathbb{N}$. Therefore, there is always an implicit word order relation $\leq_r$ in a vocabulary, that can be generated from V and c by using the order relation $\geq$ in $\mathbb{N}$[4]. Thus, a *rank* is defined as a partially-ordered set formed by a vocabulary–word order pair relation: $\langle V, \leq_r \rangle$.

Table 1 summarises some measures of the extracted vocabularies, where V_{inter} denotes the intersection of V_{ep} and V_{genia}. Notice that V_{inter} contains considerably less entries than the smallest vocabulary (V_{genia}). This shows to what extent both types of text differ and how important it is to use the correct techniques when working with domain-specific data in empirical approaches. The table also shows the number of hapax legomena (i.e. words that occur only once) in each corpus, and in this aspect both corpora are similar[5]. It also shows how sparseness affects language, since a vocabulary that is 400% bigger has only 5% less hapax legomena.

n-gram	genia	ep	google	yahoo
642	1	4	8090K	220M
African	2	2028	15400K	916M
fatty	16	22	2550K	59700K
medicine	4	643	21900K	934M
Mac	15	3	34500K	1910M
SH2	27	1	113K	3270K
advances	4	646	6200K	173M
thereby	29	2370	8210K	145M

Table 2: Distribution of some words in V_{inter}.

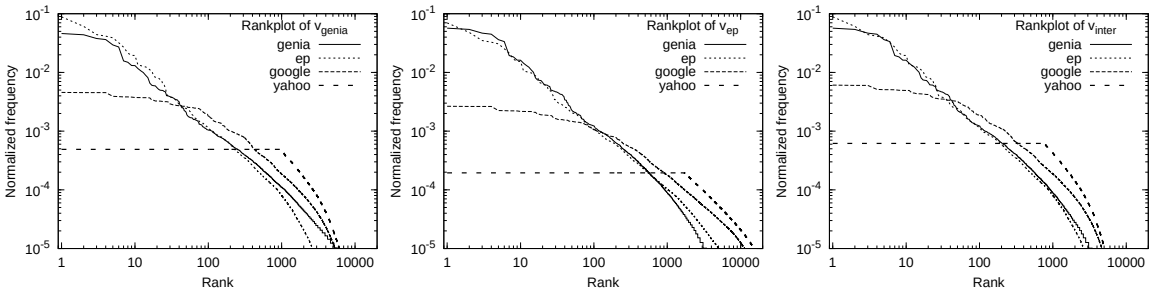
For each entry in each vocabulary, we obtained a count estimated from four different n-gram count sources: *ep*, *genia*, Google as a corpus (*google*) and Yahoo! as a corpus (*yahoo*). The latter were configured to return only results for pages in English. Table 2 shows an example of entries extracted from V_{inter}. Notice that there are no zeroes in columns *genia* and *ep*, since this vocabulary only contains words that occur at least once in these corpora. Also, some words like *Mac* and *SH2*, that are probably specialised terms, occur more in *genia* than in *ep* even if the latter is more than 80 times larger than the former.

4.3 Rank analyses

For each vocabulary, we want to estimate how similar the ranks generated by each of the four count sources are. Figure 1 shows the rank position (x) against the frequency (y) of words in V_{genia}, V_{ep} and V_{inter}, where each plotted point represents a rank position according to corpus fre-

[4] That is, $\forall v_1, v_2 \in V$, suppose $c(v_1) = n_1$ and $c(v_2) = n_2$, then $v_1 \leq_r v_2$ if and only if $n_1 \geq n_2$.

[5] The percentual difference in the proportion of hapax legomena can be explained by the fact that *genia* is much smaller than *ep*.



(a) Rank plot of V_{genia}. (b) Rank plot of V_{ep}. (c) Rank plot of V_{inter}.

Figure 1: Plot of normalised frequencies of vocabularies according to rank positions, log-log scale.

quencies and may correspond to several different words.[6] The four sources have similar shaped curves for each of the three vocabularies: *ep* and *genia* could be reasonably approximated by a linear regression curve (in the log-log domain). *google* and *yahoo* present Zipfian curves for low frequency ranges but have a flat line for higher frequencies, and the phenomenon seems consistent in all vocabularies and more intense on *yahoo*. This is related to the problem discussed in section 3 which is that *web*-based frequencies are not accurate to model common words because *web* counts correspond to page counts and not to word counts, and that a common word will probably appear dozens of times in a single page. Nonetheless, *google* seems more robust to this effect, and indeed *yahoo* returns exactly the same value (roughly 2 billion pages) for a large number of common words, producing the perfectly straight line in the rank plots. Moreover, the problem seems less serious in V_{inter}, but this could be due to its much smaller size. These results show that *google* is incapable of distinguishing among the top-100 words while *yahoo* is incapable of distinguishing among the top-1000 words, and this can be a serious drawback for *web*-based counts both in general-purpose and specialised NLP tasks.

The curves agree in a large portion of the frequency range, and the only interval in which *genia* and *ep* disagree is in lower frequencies (shown in the bottom right corner). This happens because general-purpose *ep* frequencies are much less accurate to model the specialised *genia* vocabulary, specially in low frequency ranges when sparseness becomes more marked (figure 1(a)), and vice-versa (figure 1(b)). This effect is minimised in figure 1(c), corresponding to V_{inter}.

Although both vocabularies present the same word frequency distributions, it does not mean that their ranks are similar for the four count sources. Tables 3 and 4 show the correlation scores for the compared count sources and for the two vocabularies, using Kendall's τ. The τ correlation index estimates the probability that a word pair in a given rank has the same *respective* position in another rank, in spite of the distance between the words[7].

In the two vocabularies, correlation is low, which indicates that the ranks tend to order words differently even if there are some similarities in terms of the shape of the frequency distribution. When we compare *genia* with *google* and with *yahoo*, we observe that *yahoo* is slightly less correlated with *genia* than *google*, probably because of its uniform count estimates for frequent words. However, both seem to be more similar to *genia* than *ep*.

A comparison of *ep* with *google* and with *yahoo* shows that *web* frequencies are much more similar to a general-purpose count source like *ep* than to a specialised source like *genia*. Additionally, both *yahoo* and *google* seem equally correlated to *ep*.

[6] Given the Zipfian behaviour of word probability distributions, a log-log scale was used to plot the curves.

[7] For all correlation values, $p < 0.001$ for the alternative hypothesis that τ is greater than 0.

	V_{genia}	V_{genia} *top*	V_{genia} *middle*	V_{genia} *bottom*
genia-ep	0.26	0.24	0.13	0.06
genia-google	0.28	0.24	0.18	0.09
genia-yahoo	0.27	0.22	0.17	0.09
ep-google	0.57	0.68	0.53	0.49
ep-yahoo	0.57	0.68	0.53	0.49
google-yahoo	0.90	0.90	0.89	0.89

Table 3: Kendall's τ for count sources in V_{genia}.

	V_{ep}	V_{ep} *top*	V_{ep} *middle*	V_{ep} *bottom*
genia-ep	0.26	0.36	0.07	0.04
genia-google	0.27	0.39	0.15	0.12
genia-yahoo	0.24	0.35	0.12	0.10
ep-google	0.40	0.45	0.22	0.09
ep-yahoo	0.38	0.44	0.20	0.08
google-yahoo	0.86	0.89	0.84	0.83

Table 4: Kendall's τ for count sources in V_{ep}.

Surprisingly, this correlation is higher for V_{genia} than for V_{ep}, as *web* frequencies and *ep* frequencies are more similar for a specialised vocabulary than for a general-purpose vocabulary. This could mean that the three perform similarly (poorly) at estimating frequencies for the biomedical vocabulary (V_{genia}) whereas they differ considerably at estimating general-purpose frequencies.

The correlation of the rank (first column) is also decomposed into the correlation for *top* words (more than 10 occurrences), *middle* words (10 to 3 occurrences) and *bottom* words (2 and 1 occurrences). Except for the pair *google-yahoo*, the correlation is much higher in the top portion of the vocabulary and is close to zero in the long tail. In spite of the logarithmic scale of the graphics in figure 1, that show the largest difference in the top part, the bottom part is actually the most irregular. The only exception is *ep* compared with the *web* count sources in V_{genia}: these two pairs do not present the high variability of the other compared pairs, and this means that using *ep* counts (general-purpose) to estimate *genia* counts (specialised) is similar to using *web* counts, independently of the position of the word in the rank. Counts from *google* and from *yahoo* are also very similar, specially if we also consider Spearman's ρ, that is very close to total correlation. Web ranks are also more similar for a specialised vocabulary than for a general-purpose one, providing further evidence for the hypothesis that the higher correlation is a consequence of both sources being poor frequency estimators. That is, for a given vocabulary, when *web* count sources are good estimators, they will be more distinct (e.g. having less zero frequencies).

5 Combining corpora frequencies

In our second experiment, the goal is to propose and to evaluate techniques for the combination of *n*-gram counts from heterogeneous sources. Therefore, we will use the insights about the vocabulary differences presented in the previous section. In this evaluation, we measure the impact of the suggested techniques in the identification of noun–noun compounds in corpora. Noun compounds are very frequent in general-purpose and specialised texts (e.g. *bus stop*, *European Union* and *gene activation*). We extract them automatically from *ep* and from *genia* using a standard method based on POS patterns and association measures (Evert and Krenn, 2005; Pecina, 2008; Ramisch et al., 2010).

5.1 Experimental setup

The evaluation task consists of, given a corpus of N words, extract all occurrences of adjacent pairs of nouns[8] and then rank them using a standard statistical measure that estimates the association strength between the two nouns. Analogously to the formalism adopted in section 4.2, we assume that, for each corpus, we generate a set NN containing n-grams $v_{1\ldots n} \in NN$[9] for which we obtain n-gram counts from four sources. The elements in NN are generated by comparing the POS pattern *noun–noun* against all the bigrams in the corpus and keeping only those pairs of adjacent words that match the pattern. The calculation of the association measure, considering a bigram $v_1 v_2$, is based on a contingency table which cells

[8] We ignore other types of compounds, e.g. adjective–noun pairs.

[9] We abbreviate a sequence $v_1 \ldots v_n$ as $v_{1\ldots n}$.

contain all possible outcomes $a_1a_2, a_i \in \{v_i, \neg v_i\}$. For *web*-based counts, we corrected up to 2% of them by forcing the frequency of a unigram to be at least equal to the frequency of the bigram in which it occurs. Such inconsistencies are incompatible with statistical approaches based on contingency table, as discussed in section 2.

The log-likelihood association measure (*LL*, alternatively called *expected mutual information*), estimates the difference between the observed table and the expected table under the assumption of independent events, where $E(a_1 \ldots a_n) = \frac{\prod_{i=1}^{n} c(a_i)}{N^{n-1}}$ is calculated using maximum likelihood:

$$LL(v_1v_2) = \sum_{a_1a_2} c(a_1a_2) \times \log_2 \frac{c(a_1a_2)}{E(a_1a_2)}$$

The evaluation of the *NN* lists is performed automatically with the help of existing noun compound dictionaries. The general-purpose gold standard, used to evaluate NN_{ep}, is composed of bigram noun compounds extracted from several resources: 6,212 entries from the Cambridge International Dictionary of English, 22,981 from Wordnet and 2,849 from the data sets of MWE 2008[10]. Those were merged into a single general-purpose gold standard that contains 28,622 bigram noun compounds. The specialised gold standard, used to evaluate NN_{genia}, is composed of 7,441 bigrams extracted from constituent annotation of the *genia* corpus with respect to concepts in the Genia ontology (Kim et al., 2006).

True positives (TPs) are the *n*-grams of *NN* that are contained in the respective gold standard, while *n*-grams that do not appear in the gold standard are considered false positives[11]. While this is a simplification that underestimates the performance of the method, it is appropriate for the purpose of this evaluation because we compare only the *mean average precision* (MAP) between two *NN* ranks, in order to verify whether improvements obtained by the combined frequencies are significant. Additionaly, MWEs are complex linguistic phenomena, and their annotation, specially in a domain corpus, is a difficult task that reaches low agreement rates, sometimes even for expert native speakers. Therefore, not only for theoretical reasons but also for practical reasons, we adopted an automatic evaluation procedure rathern than annotating the top candidates in the lists by hand.

Since the log-likelihood measure is a function that assigns a real value to each *n*-gram, there is a rank relation $\leq_r$ that will be used to calculate MAP as follows:

$$MAP(NN, \leq_r) = \frac{\sum_{v_{1\ldots n} \in NN} P(v_{1\ldots n}) \times p(v_{1\ldots n})}{|\text{TPs in } NN|},$$

where $p = 1$ if $v_{1\ldots n}$ is a TP, 0 else, and the precision $P(v_{1\ldots n})$ of a given *n*-gram corresponds to the number of TPs before $v_{1\ldots n}$ in $\langle NN, \leq_r \rangle$ over the total number of *n*-grams before $v_{1\ldots n}$ in $\langle NN, \leq_r \rangle$.

5.2 Combination heuristics

From the initial list of 176,552 lemmatised *n*-grams in NN_{ep} and 14,594 in NN_{genia}, we filtered out all hapax legomena in order to remove noise and avoid useless computations. Then, we counted the occurrences of v_1, v_2 and v_1v_2 in our four sources, and those were used to calculate the four *LL* values of *n*-grams in both lists. We also propose three heuristics to combine a set of m count sources c_1 through c_m into a single count source c_{comb}:

$$c_{comb}(v_{1\ldots n}) = \sum_{i=1}^{m} w_i(v_{1\ldots n}) \times c_i(v_{1\ldots n}),$$

where $w(v_{1\ldots n})$ is a function that assigns a weight between 0 and 1 for each count source according to the *n*-gram $v_{1\ldots n}$. Three different functions were used in our experiments: *uniform* linear interpolation assumes a constant and uniform weight $w(v_{1\ldots n}) = 1/m$ for all *n*-grams; *proportional* linear interpolation assumes a constant weight $w_i(v_{1\ldots n}) = ((\sum_{j=1}^{m} N_j) - N_i)/\sum_{j=1}^{m} N_j$ that is proportional to the inverse size of the corpus; and *back-off* uses the uniform interpolation of *web* frequencies whenever the *n*-gram count in the original corpus falls below a threshold (empirically defined as $\log_2(N/100,000)$).

[10] 420 entries provided by Timothy Baldwin, 2,169 entries provided by Su Nam Kim and 250 entries provided by Preslav Nakov, freely available at `http://multiword.sf.net/`

[11] In fact, nothing can be said about an *n*-gram that is not in a (limited-coverage) dictionary, further manual annotation would be necessary to asses its relevance.

MAP of rank	NN_{genia}	NN_{ep}
LL_{genia}	0.4400	0.0462
LL_{ep}	0.4351	0.0371
LL_{google}	0.4297	0.0532
LL_{yahoo}	0.4209	0.0508
$LL_{uniform}$	0.4254	0.0508
$LL_{proportional}$	0.4262	0.0520
$LL_{backoff}$	0.3719	0.0370

Table 5: Performance of compound extraction.

Table 5 shows that the performance of *backoff* is below all other techniques for both vocabularies, thus excluding it as a successful combination heuristic. The large difference between MAP scores for NN_{ep} and for NN_{genia} is explained by the relative size of the gold standards: while the general-purpose reference accounts for 16% of the size of the NN_{ep} set, the specialised reference has as many entries as 50% of NN_{genia}. Moreover, the former was created by joining heterogeneous resources while the latter was compiled by human annotators from the Genia corpus itself. The goal of our evaluation, however, is not to compare the difficulty of each task, but to compare the combination heuristics presented in each row of the table.

The best MAP for NN_{genia} was obtained with *genia*, that significantly outperforms all other sources except *ep*[12]. On the other hand, the use of *web*-based or interpolated counts in extracting specialised noun–noun compounds does not improve the performance of results based on sparse but reliable counts drawn from well-formed corpora. Nonetheless, the performance of *ep* in specialised extraction is surprising and could only be explained by some overlap between the corpora. Moreover, the interpolated counts are not significantly different from *google* counts, even if this corpus should have the weakest weight in *proportional* interpolation.

General-purpose compound extraction, however, benefits from the counts drawn from large corpora as *google* and *yahoo*. Indeed, the former significantly outperforms all other count sources, closely followed by *proportional* counts. In both vocabularies, *proportional* interpolation performs very similar to the best count source, but, strangely enough, it still does not outperform *google*. Further data inspection would be needed to explain these results for the interpolated combination and to try to shed some light on the reason why the *backoff* method performs so poorly.

[12]Significance was assessed through a standard one-tailed t test for equal sample sizes and variances, $\alpha = 0.005$.

6 Future perspectives

In this work, we presented a detailed evaluation of the use of *web* frequencies as estimators of corpus frequencies in general-purpose and specialised tasks, discussing some important aspects of corpus-based versus *web*-based *n*-gram frequencies. The results indicate that they are not only very distinct but they are so in different ways. The importance of domain-specific data for modelling a specialised vocabulary is discussed in terms of using *ep* to get V_{genia} counts. Furthermore, the *web* corpora were more similar to *genia* than to *ep*, which can be explained by the fact that "similar" is different from "good", i.e. they might be equally bad in modelling *genia* while they are distinctly better for *ep*.

We also proposed heuristics to combine count sources inspired by standard interpolation and back-off techniques. Results show that we cannot use *web*-based or combined counts to identify specialised noun compounds, since they do not help minimise data sparseness. However, general-purpose extraction is improved with the use of *web* counts instead of counts drawn from standard corpora.

Future work includes extending this research to other languages and domains in order to estimate how much of these results depend on the corpora sizes. Moreover, as current interpolation techniques usually combine two corpora, weights are estimated in a more or less ad hoc procedure (Lapata and Keller, 2005). Interpolating several corpora would need a more controlled learning technique to obtain optimal weights for each frequency function. Additionally, the evaluation shows that corpora perform differently according to the frequency range. This insight could be used to define weight functions for interpolation.

Acknowledgements

This research was partly supported by CNPq (Projects 479824/2009-6 and 309569/2009-5), FINEP and SEBRAE (COMUNICA project FINEP/SEBRAE 1194/07). Special thanks to Flávio Brun for his thorough work as volunteer proofreader.

References

Baayen, R. Harald. 2001. *Word Frequency Distributions*, volume 18 of *Text, Speech and Language Technology*. Springer.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2009. Web-scale N-gram models for lexical disambiguation. In Boutilier, Craig, editor, *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 1507–1512, Pasadena, CA, USA, July.

Evert, Stefan and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language Special issue on Multiword Expression*, 19(4):450–466.

Grefenstette, Gregory. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the Twenty-First International Conference on Translating and the Computer*, London, UK, November. ASLIB.

Keller, Frank and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics Special Issue on the Web as Corpus*, 29(3):459–484.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics Special Issue on the Web as Corpus*, 29(3):333–347.

Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics*, 33(1):147–151.

Kim, Jin-Dong, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2006. GENIA ontology. Technical report, Tsujii Laboratory, University of Tokyo.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit(MT Summit 2005)*, Phuket, Thailand, September. Asian-Pacific Association for Machine Translation.

Lapata, Mirella and Frank Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):1–31.

Nakov, Preslav. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, EECS Department, University of California, Berkeley, CA, USA.

Nicholson, Jeremy and Timothy Baldwin. 2006. Interpretation of compound nominalisations using corpus and web statistics. In Moirón, Begoña Villada, Aline Villavicencio, Diana McCarthy, Stefan Evert, and Suzanne Stevenson, editors, *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (MWE 2006)*, pages 54–61, Sidney, Australia, July. Association for Computational Linguistics.

Ohta, Tomoko, Yuka Tateishi, and Jin-Dong Kim. 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second Human Language Technology Conference (HLT 2002)*, pages 82–86, San Diego, CA, USA, March. Morgan Kaufmann Publishers.

Pecina, Pavel. 2008. Reference data for czech collocation extraction. In Gregoire, Nicole, Stefan Evert, and Brigitte Krenn, editors, *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 11–14, Marrakech, Morocco, June.

Ramisch, Carlos, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, May. European Language Resources Association.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Eisner, Jason, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech Republic, June. Association for Computational Linguistics.

Yamamoto, Mikio and Kenneth W. Church. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30.