Extracting an Equivalent Corpus from a Larger Corpus With Less Incidence on Perplexity

Etienne Denoual, Yves Lepage
ATR - Spoken Language Translation Research Laboratories

Abstract

It has become common to hear that the quality of a language model directly depends upon the amount of data used to construct it. Using larger and larger training-sets usually has the major interest of reducing model perplexity (or entropy). In the following work we address the problem of reducing a corpus in size while maintaining its linguistic productivity. We show how to extract an equivalent corpus from a larger one with little incidence on perplexity. The extracted corpus is an analogical base-set from which we can generate back the sentences of the original corpus.

Introduction

Natural language processing systems using training models usually require a fairly large amount of data to perform efficiently. It has become common to hear that the quality of a language model directly depends upon the amount of data that was used to construct it ("There is no better data than more data" [Banko 2001]). Consequently we tend to use larger and larger training-sets, a practice which has the major interest of decreasing model perplexity (or entropy). Here we turn our attention to cross-entropy and its derived perplexity (derived from entropy) in terms of N-grams of characters. The cross-entropy $H(p)$ of an N-gram model p(N-history..character i) on a test corpus $T = \{s_1, .., s_Q\}$ of Q sentences, with $s_i = \{c_1^i..c_{Q_{s_i}}^i\}$ a sentence of Q_{s_i} characters is:

$$H(p) = -\frac{1}{\sum_{i=1}^{Q} Q_{s_i}} \sum_{i=1}^{Q} [\sum_{j=1}^{Q_{s_i}} \log p_j^i] \qquad (1)$$

where $p_i = p(c_j^i | c_{j-N+1}^i..c_{j-1}^i)$.

1 Extracting an equivalent corpus

1.1 Equivalence by analogical productivity

Our concern is to find a more general definition of equivalence between corpora of different sizes that makes sense on some linguistic level. Intuitively the productivity of a corpus is related to its representativity of a given domain.

To this end, we shall use the notion of analogy as found in neogrammarian and structuralist linguistics, and partially formalised in [Lepage 96]. It has been shown that analogy can be used to extract subsets from a corpus that would be representative of the entire original one [Lepage 97]. In the following work, we show that this particular reduction in corpus size has little incidence on the perplexity (or entropy) of N-gram character models.
The property of analogy that is used in this work is that for four given sentences for which the analogical relation holds, any one sentence can be discarded as it can be regenerated from the three remaining ones. This comes from the fact that conversely, from a triple of given sentences it may be possible to generate a fourth sentence, i.e. to solve an analogical equation as shown on Figure 1.

1.2 On the fly Base-set Extraction

A base-set is a set of sentences from which we can at least regenerate the sentences deleted in the process of extraction and with no sentence in analogical relation, i.e. we can regenerate at least the original corpus. We call such a set, an analogical base-set of the corpus and in the sequel, a base for short as this set is an independent set (i.e. no sentence from the base can be generated by analogy using other sentences from the base) and as said above, a generator set for the corpus.

To construct an analogical base-set from a given corpus, we use an "on the fly" extraction algorithm previously defined in [Lepage 96], which is in fact a greedy algorithm: the corpus is processed from the first sentence to the last one. The current sentence is added to the base-set if and only if no analogy holds with any triple of previous sentences from the base-set. This algorithm ensures that an independent set is built that is also a generator.

2 Experiments

2.1 Data

We conducted a series of experiments using the English C-STAR[1] part of an aligned multi-lingual corpus, the Basic Traveller's Expressions Corpus (usu-

[1] See http://www.c-star.org .

I'd like to have seafood. : *I'd like to have Chinese food.* = *Do you like seafood?* : x
$\Rightarrow$ x = *Do you like Chinese food?*

Figure 1: Example of solving an analogical equation.

ally referred to as BTEC), a collection of sentences from the tourism and travel domain.

	Average	Std.deviation
Words/line	5.94	3.25
Characters/line	31.15	17.02

Figure 2: Words and Character per line average and standard deviation for the BTEC C-STAR corpus.

Altogether our data contain 162,318 sentences. Experiments will be performed on incremental sets of sentences from this corpus up to 142,318 sentences, and are carried out on characters (as opposed to words). The 20,000 remaining sentences are not used for training but as a test corpus to compute cross-entropy and the derived test-set perplexity.

2.2 Base-set size

The sizes of the obtained analogical base-sets are shown in Figure 3. The order of sentences is the one of the BTEC[2]. We obtained a reduction of 15.16% in the size of the entire corpus.

2.3 Base-set perplexity

For every (corpus,base) pair the corresponding perplexities are computed and shown on Figure 4 for 3-gram, 5-gram and 7-gram models (as the average size in terms of characters is around 5.24). For all three models, base-set perplexities stay close from the ones of the original corpora as size increases. Usually when a corpus is reduced in size by keeping only some of its sentences, perplexity increases. This is shown by computing the perplexity of three corpora of sentences randomly taken from the original corpus, equivalent in size to each obtained base-set, which we then averaged. For all sizes and models, base-set perplexity is inferior to the average of the corpora of equivalent size.

Here we could reduce corpus size, all the while maintaining perplexity as close as we could to the one of the original corpus. We retain linguistic productivity, as the analogical base-set can be used to (re)generate – at least – the original corpus.

Figure 5 shows comparative values for the entire BTEC applied to 3-gram, 5-gram and 10-gram models.

	Corpus	Base	Variation
Size	142,318	120,745	−15.16%
3-gram	5.315	5.313	−0.03%
5-gram	2.622	2.624	+0.10%
7-gram	2.383	2.402	+0.79%

Figure 5: Change in perplexity and size for the BTEC corpus.

3 Discussion and Future works

3.1 Perplexity and the quality of models

N-gram perplexity is defined by [Bahl 77] as $Pe(p) = 2^{H(p)}$. The lower the perplexity, the more information was conveyed by the training set [Rosenfeld 94], the better the expected quality of a language model constructed from it (provided the training set is large enough).It should be mentioned that our work might differ from others on one main point: we work on characters rather than words. Perplexity on characters gives us an indication on the range of the choice a model has to face when it tries to predict the next character.

Test-set perplexity, derived from test set cross-entropy is the most commonly used evaluation method for language models, however it is harder to interpret as new constraints arise : differences between training and test data and the choice of smoothing relative to data sparseness (this may not be as important an issue as we use characters rather than words : the number of different N-grams is lower than when using words).

However, the assumption that test-set perplexity (or entropy) is a relevant model quality indicator is still to be proven. One of the known weaknesses of perplexity is that it is poorly correlated to WER [Chen 98]; this N-gram perplexity only approximates the total perplexity of a corpus, which is logarithmically linked to the total entropy, itself having shown to have a much higher correlation with WER. It should also be pointed out that as yet no work, to our knowledge, has investigated a relationship between perplexity and character error rate. The question of whether test-set perplexity (or cross-entropy) on characters is a good indicator of model quality therefore remains open.

3.2 Generation by analogy

Our work has shown that when faced with a corpus size reduction task, one should choose an analogical

[2] The order of sentences should not significantly influence the sizes of the obtained base-sets as was observed in [Lepage 97].

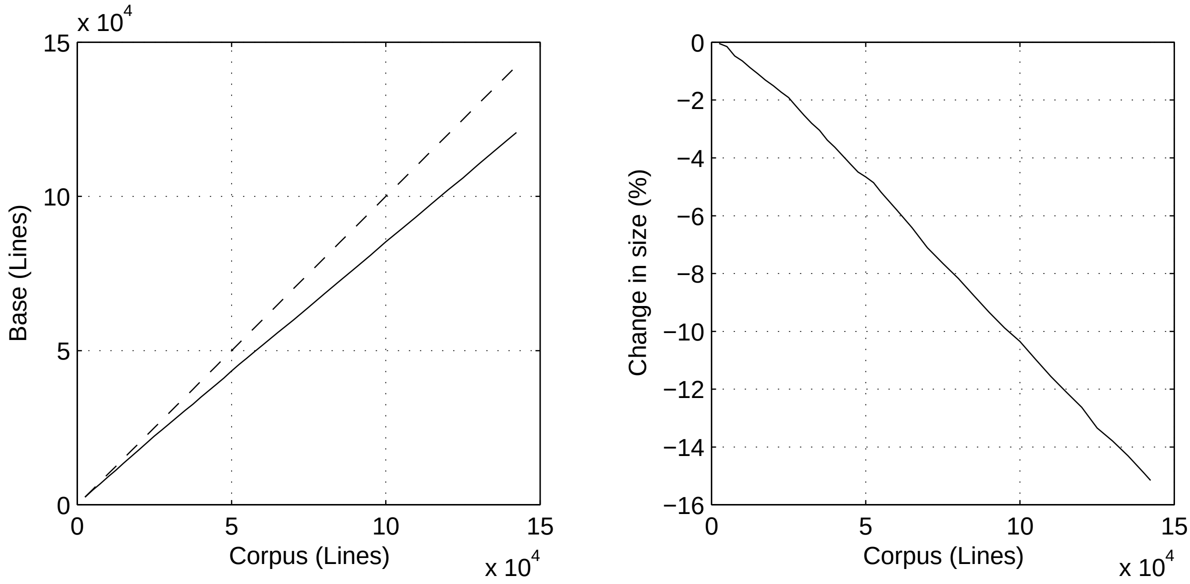

Figure 3: Size and decrease in size when extracting a base-set.

base-set of this corpus rather than randomly eliminating sentences, because the increase in perplexity will then be minimal (for low-order N-grams it will even decrease).

The next step is obviously to try and reverse the process: unlike the BTEC which is made of short to middle size sentences with a great proportion of non-unique occurences, other available corpora have greater occurence complexities. If we make the assumption that such corpora tend to be independent sets for analogy – i.e. few analogies can be extracted from such corpora – we should try to generate from them and observe variations in size and entropy. Should the process prove to be reversed correctly we could validate our assumption and be able to automatically generate larger sets with minimal perplexities.

Conclusion

By extracting an analogical base-set from a corpus of $142,318$ sentences, we have both reduced its size (-15.16%) and slightly increased its perplexity (+2.56% for 10-grams). If we make the assumption that a given corpus of large complex sentences (unlike the one used here) is an independent set for analogy, it may be possible to reverse the process – i.e. generate from that set – and increase corpus size keeping perplexity minimal. This should be investigated in future works.

4 Acknowledgements

This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

[Banko 2001] Michele Banko and Eric Brill
Scaling to Very Very Large Corpora for Natural Language Disambiguation.
Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), Toulouse,France.

[Bahl 77] Lalit R. Bahl, Jim K. Baker, Frederick Jelinek, and Robert L. Mercer
Perplexity - a measure of the difficulty of speech recognition tasks.
Program of the 94th Meeting of the Acoustical Society of America J. Acoust. Soc. Am., 62:S63, 1977. Suppl. no. 1.

[Rosenfeld 94] Ronald Rosenfeld
Adaptive Statistical Language Modeling: a Maximum Entropy Approach.
PhD Thesis (1994)

[Chen 98] Stanley Chen, Douglas Beeferman, and Ronald Rosenfeld
Evaluation Metrics For Language Models.
DARPA Broadcast News Transcription and Understanding Workshop (1998)

[Lepage 97] Yves Lepage
Corpus Contraction by Sentence Extraction Using Analogy.
Proceedings of NLPRS-97, Phuket, December 1997, pp 457-462.

[Lepage 96] Yves Lepage, Shin-Ichi Ando
Saussurian analogy: a theoretical account and its application
Proceedings of COLING-96, Copenhagen, August 1996, vol. 2, pp. 717-722.

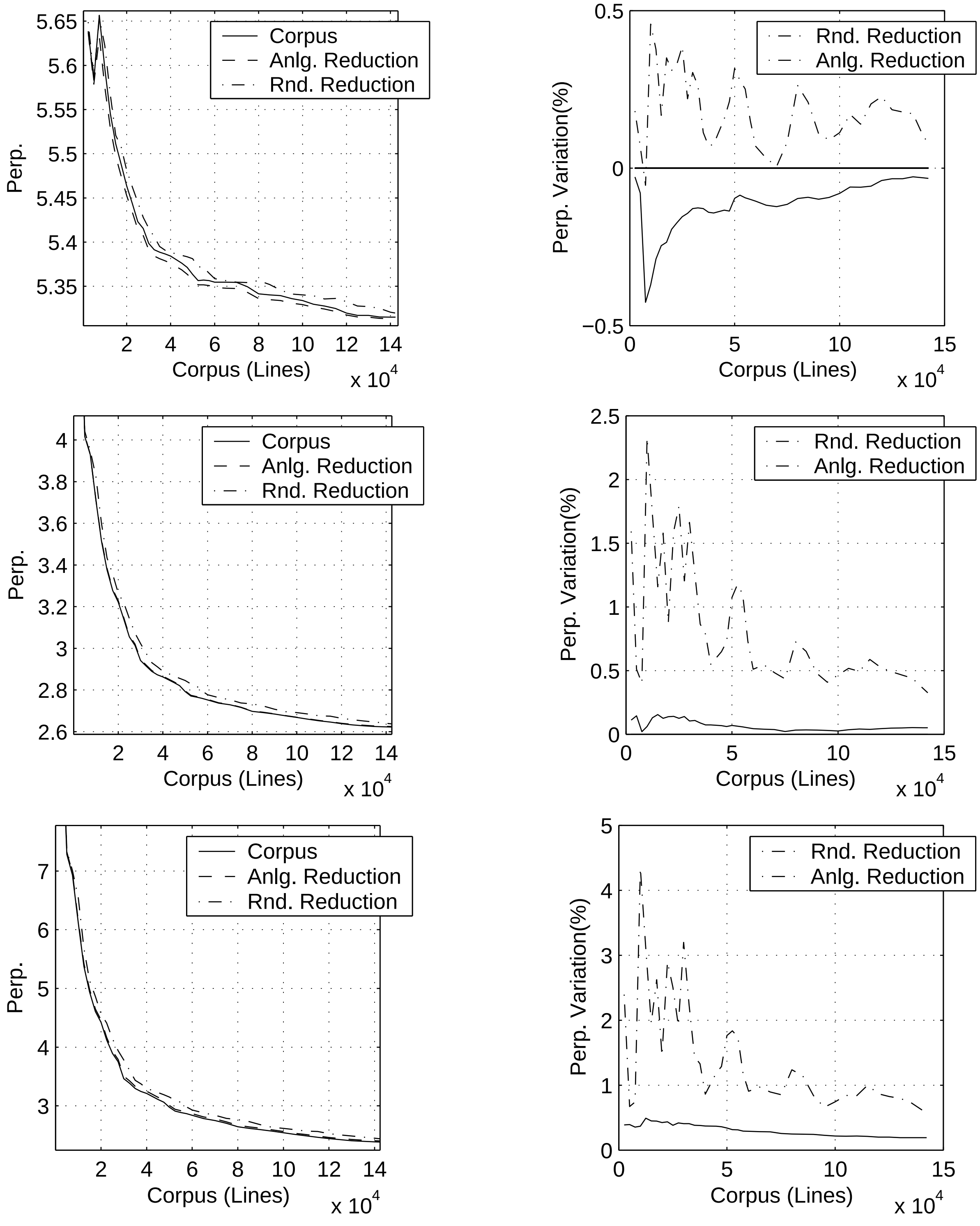

Figure 4: Perplexity and change in perplexity for 3-gram, 5-gram and 7-gram models respectively.