The influence of data homogeneity on NLP system performance

Etienne Denoual
ATR Spoken Language Communication Research Labs,
2-2-2 Keihanna Science City, Kyoto 619-0288, Japan
Laboratoire CLIPS - GETA - IMAG, Université Joseph Fourier, Grenoble, France
etienne.denoual@atr.jp

Abstract

In this work we study the influence of corpus homogeneity on corpus-based NLP system performance. Experiments are performed on both stochastic language models and an EBMT system translating from Japanese to English with a large bicorpus, in order to reassess the assumption that using only homogeneous data tends to make system performance go up. We describe a method to represent corpus homogeneity as a distribution of similarity coefficients based on a cross-entropic measure investigated in previous works. We show that beyond minimal sizes of training data the excessive elimination of heterogeneous data proves prejudicial in terms of both perplexity and translation quality : excessively restricting the training data to a particular domain may be prejudicial in terms of In-Domain system performance, and that heterogeneous, Out-of-Domain data may in fact contribute to better sytem performance.

1 Introduction

Homogeneity of large corpora is still an unclear notion. In this study we make a link between the notions of similarity and homogeneity : a large corpus is made of sets of documents to which may be assigned a score in similarity defined by cross-entropic measures (similarity is implicitly expressed in the data). The distribution of the similarity scores of such subcorpora may then be interpreted as a representation of the homogeneity of the main corpus, which can in turn be used to perform corpus adaptation to tune a corpus based NLP system to a particular domain.

(Cavaglià 2002) makes the assumption that a corpus based NLP system generally yields better results with homogeneous training data rather than heterogeneous, and experiments on a text classifier system (Rainbow[1]), to mixed conclusions. We reassess this assumption by experimenting on language model perplexity, and on an EBMT system translating from Japanese to English.

2 A framework for corpus homogeneity

2.1 Previous work on corpus similarity and homogeneity

A range of measures for corpus similarity have been put forward in past literature : (Kilgarriff and Rose 98; Kilgarriff 2001) investigated on the similarity of corpora and compared “Known Similarity Corpora” (KSC) using perplexity and cross-entropy on words, word frequency measures, and a χ^2-test which they found to be the most robust. However(as acknowledged in (Kilgarriff and Rose 98)), using KSC requires that the two corpora chosen for comparison are sufficiently similar that the most frequent lexemes in them almost perfectly overlap. However (Liebscher 2003) showed by comparing frequency counts of different large Google Group

[1] See http://www.cs.cmu.edu/mccallum/bow .

corpora that it is not usually the case.

Measuring homogeneity by counting word / lexeme frequencies introduces additional difficulties as it assumes that the word is an obvious, well-defined unit, which is not the case in the Chinese (Sproat and Emerson 2003) or Japanese language (Matsumoto et al., 2002), for instance, where word segmentation is not trivial.

(Denoual 2004) showed that similarity between corpora could be quantified with a coefficient based on the cross-entropies of probabilistic models built upon reference data. The approach needed no explicit selection of features and was language independent, as it relied on character based models (as opposed to word based models) thus bypassing the word segmentation issue and making it applicable on any electronic data.

The cross-entropy $H_T(A)$ of an N-gram model p constructed on a training corpus T, on a test corpus $A = \{s_1, .., s_Q\}$ of Q sentences with $s_i = \{c_1^i..c_{|s_i|}^i\}$ a sentence of $|s_i|$ characters is:

$$H_T(A) = \frac{\sum_{i=1}^{Q}[\sum_{j=1}^{|s_i|} -logp_j^i]}{\sum_{i=1}^{Q}|s_i|} \quad (1)$$

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

We therefore define a scale of similarity between two corpora on which to rank any third given one. Two reference corpora T_1 and T_2 are selected by the user, and used as training sets to compute N-gram character models. The cross-entropies of these two reference models are estimated on a third test set T_3, and respectively named $H_{T_1}(T_3)$ and $H_{T_2}(T_3)$ as in the notation in Eq. 1. Both model cross-entropies are estimated according to the other reference , i .e ., $H_{T_1}(T_2)$ and $H_{T_1}(T_1)$, $H_{T_2}(T_1)$ and $H_{T_2}(T_2)$ so as to obtain the weights W_1 and W_2 of references T_1 and T_2 :

$$W_1 = \frac{H_{T_1}(T_3) - H_{T_1}(T_1)}{H_{T_1}(T_2) - H_{T_1}(T_1)} \quad (2)$$

and :

$$W_2 = \frac{H_{T_2}(T_3) - H_{T_2}(T_2)}{H_{T_2}(T_1) - H_{T_2}(T_2)} \quad (3)$$

After which W_1 and W_2 are assumed to be the weights of the barycentre between the user-chosen references. Thus

$$I(T_3) = \frac{W_1}{W_1 + W_2} = \frac{1}{1 + \frac{W_2}{W_1}} \quad (4)$$

is defined to be the similarity coefficient between reference sets 1 and 2, which are respectively corpus T_1 and corpus T_2 . Given the previous assumptions, $I(T_1) = 0$ and $I(T_2) = 1$; furthermore, any given corpus T_3 yields a score between the extrema $I(T_1) = 0$ and $I(T_2) = 1$

This framework may be applied to the quantification of the similarity of large corpora, by projecting them to a scale defined implicitly via the reference data selection. In this study we shall specifically focus on a scale of similarity bounded by a sublanguage of spoken conversation on the one hand, and a sublanguage of written style media on the other.

We build upon this previous work in order to represent intra-corpus homogeneity.

2.2 Representing corpus homogeneity

Corpora are collected sets of documents usually originating from various sources. Whether a corpus is homogeneous in content or not is scarcely known besides the knowledge of the nature of the sources. As homogeneity is multidimensional (see (Biber 1988) and (Biber 1995) for considerations on the dimensions in register variation for instance), one cannot trivially say that a corpus is homogeneous or heterogeneous : different sublanguages show variations that are lexical, semantic, syntactic, and structural (Kittredge and Lehrberger 1982).

In this study we wish to implicitly capture such variations by applying the previously described similarity framework to the representation of homogeneity. Coefficients of similarity may be computed for all smaller sets in a corpus, the distribution of which shall depict the homogeneity of the corpus relatively to the scale defined implicitly by the choice of the reference data.

Homogeneity as depicted here is relative to the choice of reference training data, which implicitly embrace lexical and syntactic variations in a sublanguage (which are by any means not unidimensional, as argued previously). We focus as in (Denoual 2004) on a scale of similarity bounded by a sublanguage of spoken conversation on the one hand, and a sublanguage of written style media on the other.

3 A study of the homogeneity of a large bicorpus

3.1 Data

Reference data is needed to set up a scale of similarity, and implicitly bound it.

For the sublanguage of spoken conversation we used for both English and Japanese the SLDB (Spontaneous Speech Database) corpus, a multilingual corpus of raw transcripts of dialogues described in (Nakamura et al., 1996).

For the sublanguage of written style media, we used for English a part of the Calgary[2] corpus, containing several contemporary English literature pieces[3], and for Japanese a corpus of collected articles from the Nikkei Shinbun newspaper[4].

The large multilingual corpus that is used in our study is the C-STAR[5] Japanese / English part of an aligned multilingual corpus, the Basic Traveller's Expressions Corpus (BTEC).

A prerequisite of the method is that levels of data transcriptions are strictly normalized, so that the comparison is not made on the transcription method but on the underlying signal itself.

3.2 Homogeneity in the BTEC

The BTEC is a collection of sentences originating from 197 sets (one set originating from one phrasebook) of basic travel expressions. Here we examine the distribution of the similarity coefficients assigned to its subsets.

The corpus may be segmented in a variety of manners, however we wish to proceed in two intuitive ways : firstly, by keeping the original subdivision, i .e ., one phrasebook per subset ; secondly, at the level of the sentence, i .e ., one sentence per subset . Figure 1 shows the similarity coefficient distributions for Japanese and English at the sentence and subset level, and Table 1 shows their means and standard deviations.

The difference in means and standard deviation values can be explained by the fact that all phrasebooks do not have the same size in lines[6]. The distribution of similarity coefficients at the line level, however similar to the distribution at the phrasebook level, suggests in its irregularities that it is indeed safer to use a larger unit to estimate cross-entropies. Moreover, we wish not to tamper with the integrity of the original subsets, that is to keep the integrity of phrasebook contents as much as possible.

Coefficient	Japanese	English
Phrasebook	0.330 ± 0.020	0.288 ± 0.027
Line	0.315 ± 0.118	0.313 ± 0.156

Table 1: Means $\pm$ standard deviations of the similarity coefficient distributions in Japanese and English.

On the phrasebook level, the similarity coefficient has a low correlation on both the average phrasebook length (0.178) and the average line length (0.278) (which does not make it a too "shallow" profiling method). On the other hand, correlation is high between the coefficients in Japanese and English (0.781), which is only to be expected intuitively.

4 Experiments

4.1 Method

This work wishes to reassess the assumption that, for a same amount of training data, a corpus-based NLP system performs better when its data tends to be homogeneous. Here we use the representation of homogeneity defined by the similarity coefficient scale to select data that tends to be homogeneous to an expected task. Experiments shall be performed both on randomly selected data, and on data selected according to their similarity coefficient. The closer the coefficient of the training data is to the coefficient of the expected task, the better.

We assume that the task is sufficiently represented by a set of data from the same domain as the large bicorpus used, the BTEC. Experiments are performed on a test set of 510 Japanese sentences which are not included in the ressource.

[2]The Calgary Corpus is available via anonymous ftp at ftp.cpcs.ucalgary.ca/pub/projects/text.compression.corpus .

[3]Parts are entitled book1, book2 and book3.

[4]The use of classical Japanese literature is not appropriate as (older) copyright-free works make use of a considerably different language. In order to maintain a certain homogeneity, we limit our study to contemporary language.

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

[6]The BTEC phrasebooks have an average size of 824 lines with a standard deviation in size of 594 lines.

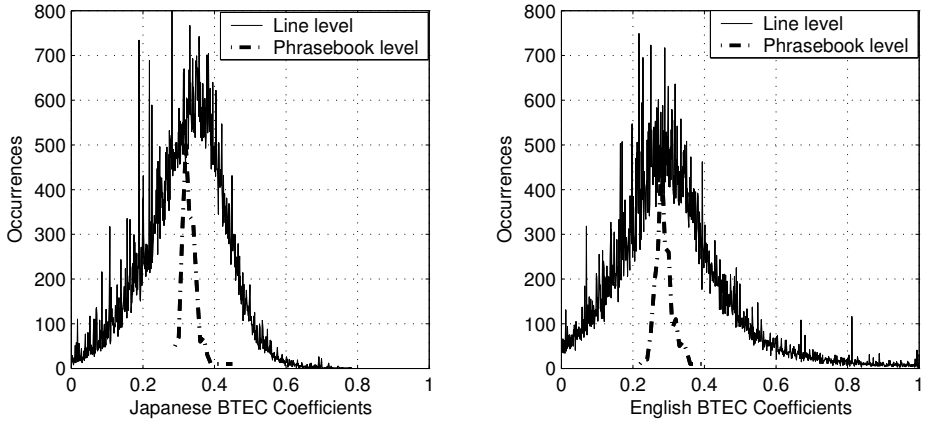

Figure 1: Distributions of similarity coefficients at the sentence level (thin line) and at the phrasebook level (thick line), respectively for Japanese and English.

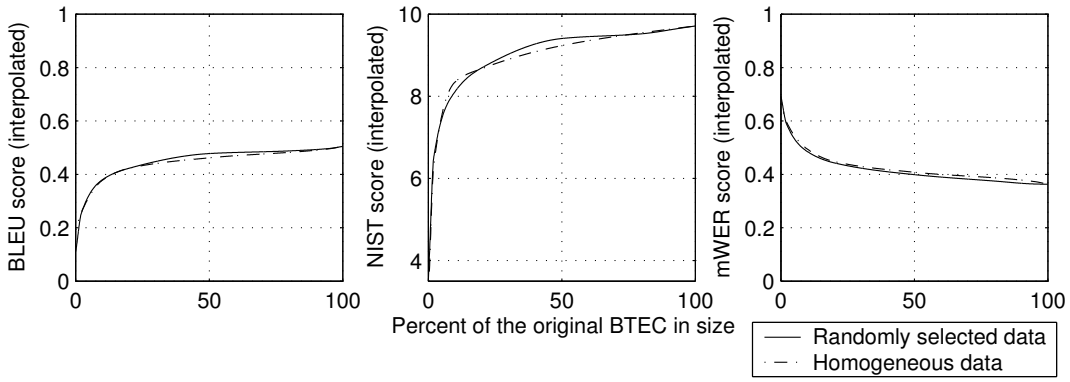

Figure 2: BLEU, NIST and mWER scores for EBMT systems built on increasing amounts of randomly chosen and homogeneous BTEC data.

These sentences shall first be used for language model perplexity estimation, then as input sentences for the EBMT system. The task is found to have a coefficient of $I_0 = 0.331$. The average coefficient for a BTEC phrasebook being 0.330, the task is found to be particularly in the domain of the ressource. We examine the influence of training data size first on language model perplexity, then on the quality of translation from Japanese to English by an example-based MT system.

4.1.1 Language model perplexity

Even if perplexity does not always yield a high correlation with NLP systems performance, it is still an indicator of language model complexity as it gives an estimate of the average branching factor in a language model. The measure is popular in the NLP community because admittedly, when perplexity decreases, the performance of systems based on stochastic models tends to increase.

We compute perplexities of character language models built on variable amounts of training data first randomly taken from the Japanese part of the BTEC, and then selected around the expected task coefficient I_0 (thresholds are determined by the amount of training data to be kept). Cross-entropies are estimated on the test set, and all estimations are performed five times for the random data selections and averaged. Figure 3 shows the character perplexity values for increasing amounts of data from 0.5% to 100% of the BTEC and interpolated. As was expected, perplexity decreases as training data increases and tends to have an asymptotic behaviour when more data is being used as training.

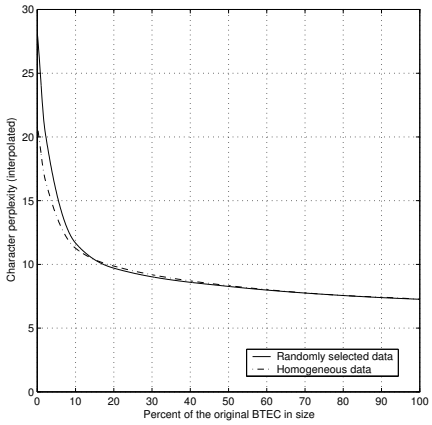

Figure 3: Perplexity of character language models built on increasing amounts of randomly chosen BTEC and homogeneous Japanese data.

While homogeneous data yield lower perplexity scores for small amounts of training data (up to 15% of the ressource - roughly 1.5 Megabytes of data), beyond this value perplexity is slightly higher than for a model trained on randomly selected data. Except for the smaller amounts of data, there seems to be no benefit in using homogeneous rather than random heterogeneous training data for model perplexity. On the contrary, excessively restricting the domain seems to yield higher model perplexities.

4.1.2 Automatic evaluation of the translation quality

In this section we experiment on a Japanese to English grammar-based EBMT system, HPATR (described in (Imamura 2001)), which parses a bicorpus with grammars for both source and target language, and translates by automatically generating transfer patterns from bilingual trees constructed on the parsed data. Not being a MT system based on stochastic methods, it is used here as a task evaluation criterion complementary to language model perplexity. Systems are likewise constructed on variable amounts of training data, and evaluated on the previous task of 510 Japanese sentences, to be translated from Japanese to English.

Because it is not feasible here to have humans judge the quality of many sets of translated data, we rely on an array of well known automatic evaluation measures to estimate translation quality :

- BLEU (Papineni et al. 2002) is the geometric mean of the n-gram precisions in the output with respect to a set of reference translations. It is bounded between 0 and 1, better scores indicate better translations, and it tends to be highly correlated with the fluency of outputs ;
- NIST (Doddington 2002) is a variant of BLEU based on the arithmetic mean of weighted n-gram precisions in the output with respect to a set of reference translations. It has a lower bound of 0, no upper bound, better scores indicate better translations, and it tends to be highly correlated with the adequacy of outputs ;
- mWER (Och 2003) or Multiple Word Error Rate is the edit distance in words between the system output and the closest reference translation in a set. It is bounded between 0 and 1, and lower scores indicate better translations.

Figure 2 shows BLEU, NIST and mWER scores for increasing amounts of data from 0.5% to 100% of the BTEC and interpolated. As was expected, MT quality increases as training data increases and tends to have an asymptotic behaviour when more data is being used in training. Here again except for the smaller amounts of data (up to 3% of the BTEC in BLEU, up to 18% in NIST and up to 2% in mWER), using the three evaluation methods, translation quality is equal or higher when using random heterogenous data. If we perform a mean comparison of the 510 paired score values assigned to sentences, for instance at 50% of training data, this difference is found to be statistically significant between BLEU, NIST, and mWER scores with confidence levels of 88.49%, 99.9%, and 73.24% respectively.

5 Discussion and future work

The contribution of this work is twofold :

We describe a method of representing homogeneity according to a cross-entropic measure of similarity to reference sublanguages, that can be used to profile language ressources. A corpus is represented by the distribution of the similarity coefficients of the smaller subsets it contains,

and atypical therefore heterogeneous data may be characterized by the lower occurrences of their values.

We further observe that marginalizing such atypical data in order to restrict the domain on which a corpus-based NLP system operates does not yield better performance, either in terms of perplexity when the system is based on stochastic language models, or in terms of objective translation quality when the system is a grammar-based EBMT system.

An objective for future work is therefore to study corpus adaptation with Out-of-Domain data. While (Cavaglià 2002) also acknowledged that for minimal sizes of training data, the best NLP system performance is reached with homogeneous ressources, we would like to know more precisely why and to what extent mixing In-Domain and Out-of-Domain data yields better accuracy. Concerning the representation of homogeneity, other experiments are needed to tackle the multidimensionality of sublanguage varieties less implicitly. We would like to consider multiple sublanguage references to untangle the dimensions of register variation in spoken and written language.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology.

References

Douglas Biber. 1988. *Variation across speech and writing.* Cambridge University Press.

Douglas Biber. 1995. *Dimensions in Register Variation.* Cambridge University Press.

Gabriela Cavaglià. 2002. *Measuring corpus homogeneity using a range of measures for inter-document distance.* Proceedings of LREC, pp. 426-431.

Etienne Denoual. 2004. *A method to quantify corpus similarity and its application to quantifying the degree of literality in a document.* Proceedings of the International Workshop on Human Language Technology, Hong Kong, pp.28-31.

George Doddington. 2002. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.* Proceedings of Human Lang. Technol. Conf. (HLT-02), pp.138-145.

Kenji Imamura. 2001. *Hierarchical Phrase Alignment Harmonized with Parsing.* Proceedings of NLPRS, pp.377-384.

Adam Kilgarriff and Tony Rose. 1998. *Measures for corpus similarity and homogeneity.* Proceedings of the 3rd conference on Empirical Methods in Natural Language Processing, Granada, Spain, pp. 46 - 52.

Adam Kilgarriff. 2001. *Comparing corpora.* International Journal of Corpus Linguistics 6:1, pp. 1-37.

Richard Kittredge and John Lehrberger. 1982. *Sublanguage. Studies of language in restricted semantic domains* Walter de Gruyter, editor.

Robert A. Liebscher. 2003. *New corpora, new tests, and new data for frequency-based corpus comparisons.* Center for Research in Language Newsletter, 15:2

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda,Kazuma Takaoka and Masayuki Asahara. 2002. *Morphological Analysis System ChaSen version 2.2.9 Manual.* Nara Institute of Science and Technology.

Atsushi Nakamura, Shoichi Matsunaga, Tohru Shimizu, Masahiro Tonomura and Yoshinori Sagisaka 1996. *Japanese speech databases for robust speech recognition.* Proceedings of the ICSLP'96, Philadelphia, PA, pp.2199-2202, Volume 4

Franz Josef Och. 2003. *Minimum Error Rate Training in Statistical Machine Translation.* Proceedings of ACL 2003, pp.160-167.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation.* Proceedings of ACL 2002, pp.311-318.

Richard Sproat and Thomas Emerson. 2003. *The First International Chinese Word Segmentation Bakeoff.* The Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan.