

The influence of example-data homogeneity on EBMT quality

Etienne Denoual

ATR Spoken Language Translation 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

1 Introduction

Homogeneity of large corpora is still a largely unclear notion. In this study we first 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, such similarity being 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. A blatant fact is that the quality of an example-based machine translation (EBMT) system will depend heavily on the training examples it is fed. Being able to tune an MT system to a specific application through a wise selection of training data is therefore a critical issue. From this viewpoint, such a representation of homogeneity may be used to perform corpus adaptation to tune an EBMT system to the particular domain, or sublanguage, of an expected task. In the following study we further describe this framework and compare it with existing methods based on computing linguistic feature frequencies.

(Cavaglia 2002) made the general assumption that a corpus-based NLP system generally yields better results with homogeneous rather than heterogeneous training data, and experimented on a text classifier system (Rainbow¹), with mixed conclusions. Not finding such an assumption completely straightforward, we reassess it by experimenting on language model perplexity, and on a grammar-based EBMT system translating from Japanese to English, in order to see if there is a real correlation between EBMT system performance and the homogeneity of the corpus of examples.

¹See <http://www.cs.cmu.edu/mccallum/bow> .

2 A framework for corpus homogeneity

2.1 Previous work on corpus similarity and homogeneity

Corpus similarity has been extensively studied in past literature, and a wide range of measures have been put forward: (Kilgarriff and Rose 98; Kilgarriff 2001) investigated the similarity and homogeneity of corpora and proceeded to compare “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)), such a comparison methodology requires that the two corpora chosen for comparison are sufficiently similar that the most frequent lexemes in them almost perfectly overlap. Whereas intuition would hint at this being true for very large corpora, (Liebscher 2003) showed by comparing frequency counts of different Google Group corpora that it is generally not the case. Furthermore, measuring homogeneity by counting word / lexeme frequencies introduces another additional difficulty: this assumes that the word is a clearly defined unit, which is not the case in the Chinese (Sproat and Emerson 2003) or Japanese language (Matsumoto et al., 2002), for instance, where there is no word segmentation.

We claim that similarity between corpora can be adequately quantified with a coefficient based on the cross-entropies of probabilistic models, built upon reference data. The approach needs no explicit selection of features and is language independent, as it relies 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 $A = \{s_1, \dots, s_Q\}$ of Q sentences with $s_i = \{c_1^i \dots c_{|s_i|}^i\}$ a sentence of $|s_i|$ characters is:

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

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

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 barycenter between the user-chosen references. Thus

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

is defined to be the similarity coefficient between reference sets 1 and 2, which are respectively corpus T_1 and corpus T_2 . Let us point out that given the previous assumptions, $I(T_1) = 0$ and $I(T_2) = 1$; furthermore, any given corpus T_3 is then awarded 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 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.

2.2 Experimental data used

To set up a scale of similarity between spoken conversation style data and written style docu-

ments, we need to select reference data which shall implicitly bound the scale.

For the sublanguage of spoken conversation we used for both English and Japanese languages 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 the English language a part of the Calgary² corpus, familiar in the data-compression field, containing several contemporary English literature pieces³, and for the Japanese language a corpus of collected articles from the Nikkei Shinbun newspaper⁴.

The large multilingual corpus that is used in our study is the C-STAR⁵ Japanese/English part of an aligned multilingual corpus, the Basic Traveller's Expressions Corpus (BTEC).

Statistical aspects for each corpus are shown in Tables 1 and 2 for English and Japanese.

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 data itself.

2.3 A comparison with other existing similarity measures

As mentioned in Section 2.1, a number of similarity measures have been investigated, which make use of linguistic feature counts such as the frequency lists of words or lexemes. Such methods assume that the word is a well-defined unit, or rely on the use of segmenters when dealing with languages in which text is not segmented into words. We wish to compare our proposed method to two measures based on feature frequency computation, which have been previously applied to English corpora in past literature: Chi Square (χ^2) and Log-likelihood (G^2). Both measures are symmetric, and compare one document to another via their feature frequency lists. The output number is interpreted as an

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

³Parts are entitled book1, book2 and book3.

⁴The 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.

⁵See <http://www.c-star.org> .

English corpora	SLDB	BTEC	Calgary
Word/Sent.	11.27±6.85	5.94±3.25	20.21±15.18
Char./Sent.	64.51±35.95	31.15±17.02	107.70±84.69
Char./Word	5.72	5.24	5.33
Total Char.	1,037K	5,026K	757K
Total Words	181.2K	964.2K	142.2K
Total Sent.	16,078	162,318	7,035

Table 1: Statistical aspects of several English corpora. (Mean ± std. dev)

Japanese corpora	SLDB	BTEC	Nikkei
Char./Stce (Mean)	32.61±22.22	14.45±7.12	44.21±28.34
Total Char.	20,806K	2,426K	2,772K
Total Sent.	84,751	162,318	253,016

Table 2: Statistical aspects of several Japanese corpora. (Mean ± std. dev)

inter-document distance.

2.3.1 Similarity measures in previous works

The Chi Square measure (χ^2), as in (Kilgarriff 2001): the number of occurrences of a feature that would be expected in each document is calculated from the frequency lists. If the sizes of documents A and B are respectively N_A and N_B , and feature w has been observed with a frequency of $o_{w,A}$ in A and $o_{w,B}$ in B , then the expected value $e_{w,A}$ is:

$$e_{w,A} = \frac{N_A(o_{w,A} + o_{w,B})}{N_A + N_B} \quad (5)$$

and likewise for $e_{w,B}$ for document B . The χ^2 value for the document pair A and B is then computed as follows:

$$\chi^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i} \quad (6)$$

with the sum over the n features.

The Log-likelihood measure (G^2): (Dunning 1993) showed that G^2 is a better approximation of the binomial distribution than χ^2 , especially for less frequent events. It was shown to work well with documents of various sizes and to allow the comparison of both frequent and rare events. G^2 is the sum of the log-likelihoods G_w^2 of all n features w :

$$G_w^2 = 2(a \log(a) + b \log(b) + c \log(c) + d \log(d))$$

$$\begin{aligned} & - (a + b) \log(a + b) - (a + c) \log(a + c) \\ & - (b + d) \log(b + d) - (c + d) \log(c + d) \\ & + (a + b + c + d) \log(a + b + c + d) \end{aligned} \quad (7)$$

	Doc.A	Doc.B
w	a	b
$\neg w$	c	d

Table 3: Contingency table for feature w in documents A and B .

a , b , c and d being defined for each feature by the contingency table given in Table 3, so that in the end:

$$G^2 = \sum_{i=1}^n G_i^2 \quad (8)$$

Both measures yield a value which is interpreted as the inter-document distance between two documents. Such distances can in turn be transposed in the view of our framework, so as to define similarity coefficients based on G^2 and χ^2 (i.e., character cross-entropy $H_T(A)$ is replaced in our framework by χ^2 or G^2 measures).

2.3.2 Evaluation

In order to compare our method with the alternative similarity coefficients based on G^2 and χ^2 , we use the method of Known Similarity Corpora (KSC) as in (Kilgarriff 2001). The comparison will be performed on Japanese, a language without clear word segmentation, so that

text data will first have to be run through an analyser when using G^2 and χ^2 distances. To allow a fair comparison, our method will be applied on raw unsegmented data. We construct three sets of KSCs with the previously described SLDB, BTEC and Nikkei corpora (See Section 2.2): slices of 10,000 words (or their equivalent in unsegmented data) are taken from each corpus and randomly rearranged so that each KSC set includes different mixes of one pair of corpora. For instance, the KSC set of SLDB and BTEC includes a subset *s10b0* containing ten slices of SLDB and zero slices of BTEC (100% SLDB, 0% BTEC), a subset *s9b1* of nine slices of SLDB and one slice of BTEC (90% SLDB, 10% BTEC), and so on. Each subset is made of ten slices and is therefore the equivalent of 100,000 words of data, on which we can produce a number of Gold Standard assertions, such as “*s10b0* should be ranked with a lower coefficient than *s9b1* because all its data comes from the corpus SLDB” (if we assume that corpora more similar to SLDB get low coefficients, and more similar to BTEC, high coefficients). Each KSC set is made of 11 subsets of 100,000 words of data. The equivalent of 500,000 words of data is left out to be used as references for distance/entropy estimation in our framework. As in (Cavaglia 2002), frequency lists include the 500 most frequent features in each document (preliminary experiments having shown that best results were achieved for 320 to 640 features).

Once KSC sets have been prepared they are scored on the three coefficients and ranked accordingly. The ranks are then compared to the Gold Standard rankings through the computation of Kappa coefficients, and Spearman rank order correlations. Results are shown in Table 4.

The KSC method has the following limitations to its validity: firstly, it does not compare different language varieties but rather mixes of the same varieties. Secondly, the size of slices may be too small to allow a fair comparison, as one corpus used in a KSC set might include highly heterogeneous parts. All three measures display very high correlations with the Gold Standard rankings. This only tends to confirm their validity as similarity indicators, at least when dealing with mixes of the same varieties of language. The best scores differ depending on the KSC sets, showing no superiority of one measure over the other two. However, our

method could be applied to Japanese data with no prior preprocessing, such as word segmentation, which makes its range of application wider than any measure relying on counting linguistic features such as words or lexemes.

2.4 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 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: 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.

Whereas the corpus may be segmented in a variety of manners, 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 sen-

Kappa	$I_{Entropy}$	I_{χ^2}	I_{G^2}
SLDB-BTEC	0.5	0.7	0.8
SLDB-Nikkei	0.9	0.7	0.7
BTEC-Nikkei	0.6	0.9	0.9

Spearman	$I_{Entropy}$	I_{χ^2}	I_{G^2}
SLDB-BTEC	0.918	0.973	0.990
SLDB-Nikkei	1.000	0.936	0.990
BTEC-Nikkei	0.982	1.000	1.000

Table 4: Kappa coefficients (ten intervals) and Spearman correlation scores of rank orders produced by similarity coefficients based on entropy, χ^2 and G^2 compared to the Gold Standard ranks.

tence and subset level, and Table 5 shows their means and standard deviations.

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

Table 5: Means \pm standard deviations of the similarity coefficient distributions in Japanese and English.

The difference in means and standard deviation values is explained by the fact that all phrasebooks do not have the same size in lines⁶. 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.

Let us point out that 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 similar amount of training data,

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

an example-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 are 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 more appropriate.

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 randomly taken from the resource (and excluded from the training set). These sentences are first 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 random selection of the test set making sure that the task is particularly in the domain of the overall resource. 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 have a high correlation with NLP system performance, it is still a valuable 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

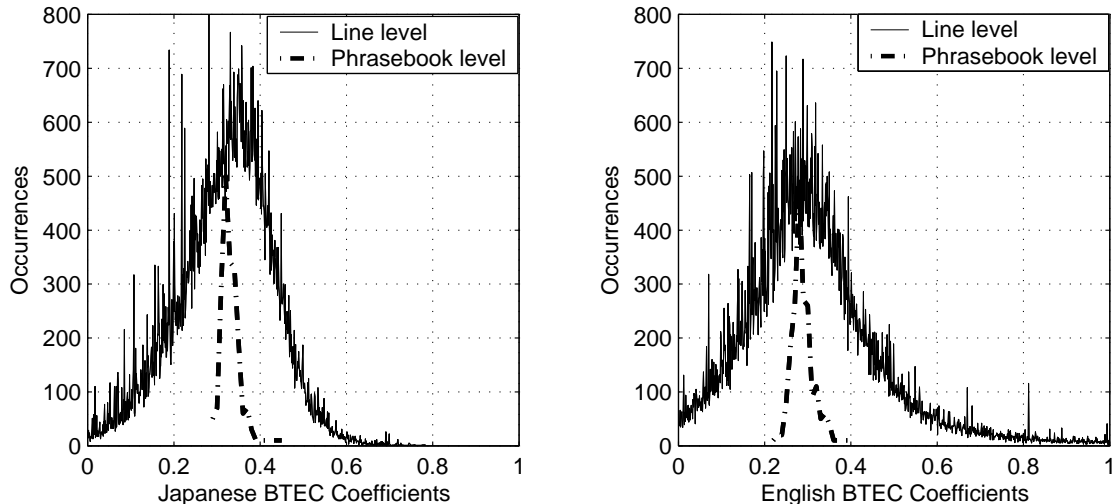


Figure 1: Distributions of similarity coefficients at the sentence level (thin line) and at the phrasebook level (thick line), respectively for Japanese and English.

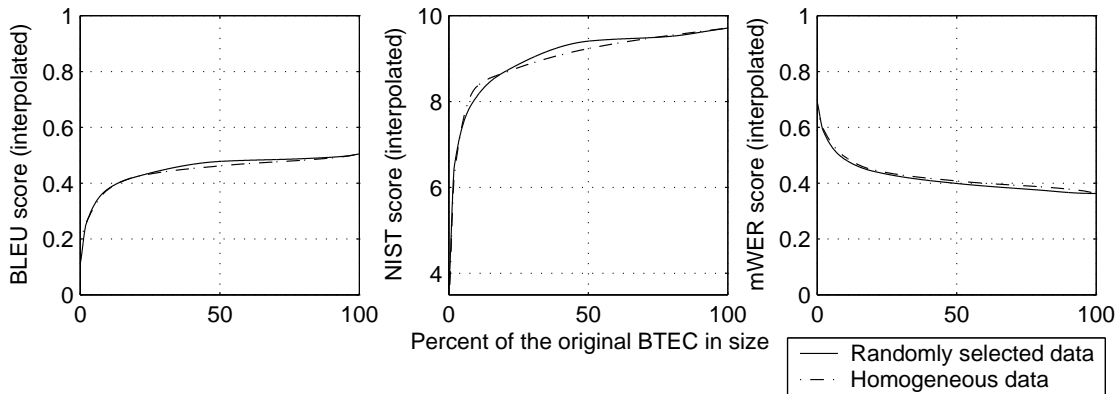


Figure 2: BLEU, NIST and mWER scores for EBMT systems built on increasing amounts of randomly chosen and homogeneous BTEC data.

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 510 sentence 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 to be expected, perplexity decreases as the amount of training data increases and tends to have an asymptotic be-

haviour when more data is being used as training.

While homogeneous data yield lower perplexity scores for small amounts of training data (up to 15% of the resource - 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 indeed 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.

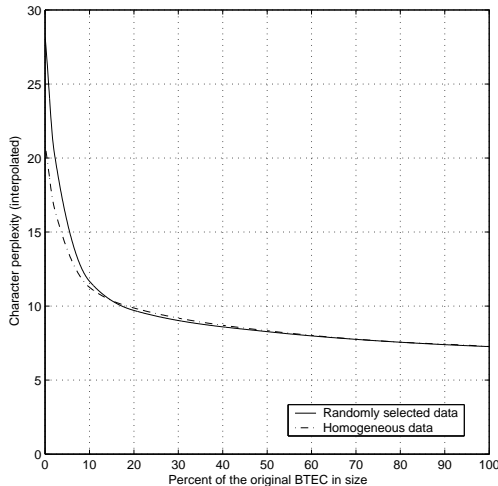


Figure 3: Perplexity of character language models built on increasing amounts of randomly chosen BTEC and homogeneous Japanese data.

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. Translation is done by automatically generating transfer patterns from bilingual trees constructed on the parsed data. Not being an MT system based on stochastic methods, it is conveniently 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 same 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, higher 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, higher 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 when using random heterogeneous data is found to be equal or higher than when using homogeneous 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 similarity to reference sublanguages through a cross-entropic measure, that can be used to profile the homogeneity of language resources. Comparing our approach to other existing similarity measures shows similar performance, while extending widely their range of application to electronic data written in languages with no clear word segmentation. A corpus may be 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 with an EBMT system.

Having observed that heterogeneous data in a resource may indeed contribute to better NLP system performance, one of our objectives for future work is 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 resources, we would like to know more precisely why and to what extent mixing In-Domain and Out-of-Domain data could yield better accuracy.

As far as the representation of homogeneity is concerned, 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.

6 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.
- 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.
- Ted Dunning. 1993. *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, 19(2):219-41.
- 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.