The character as an appropriate unit of processing for non-segmenting languages

Etienne Denoual & Yves Lepage
ATR 音声言語コミュニケーション研究所
619-0288「けいはんな学研都市」光台 2-2-2
{etienne.denoual,yves.lepage}@atr.jp

1 Introduction

Natural language processing in non-segmenting languages implies specific issues, as there is no explicit information on word or chunk boundary. This is due to the fact that a unit of word cannot be precisely defined in general linguistics. In this paper we propose the use of characters as an immediate given unit in NLP, and demonstrate a wide range of successful applications in linguistic data profiling, machine translation and its evaluation, and automatic paraphrase generation. The use of characters as a unit of processing eliminates the need of specific preprocessing tools when dealing with non-segmenting languages.

2 The word-segmentation problem

Linguists and researchers in NLP face an ever-present issue as they handle language through computers : more often than not, they handle the written form of language rather than language itself in its spoken form. However, while the written form of language is separate from the internal structure of language itself, it may not be neglected as it is very often the only given (for instance for data harvested on the web, or in large linguistic databases like those advertised and disseminated by the LDC and ELRA/ELDA). In this context, the researcher cannot avoid to ponder on the following issue : which unit is more appropriate for the processing of the written form of language?

In NLP, it has become a habit to perform computations at the level of what intuition hints at being a word in segmenting languages, a practice which has been relatively satisfactory on the side of performance, if not theoretically. Segmenting languages are languages which written form imply the use of specific separators between strings of characters, whereas in the written form of non-segmenting languages, whole sentences, paragraphs or even documents may be written continuously without any separator. For instance, in the following English sentence : *The cat eats the mouse.* one may count 5 words. Similarly, in the French sentence : *Le chat mange la souris.*, one may also count 5 words. However, in the Japanese sentence 猫が鼠を食べている。 one cannot clearly count a number of words. In this last example, we face the issue of not having any obvious word boundary implied by the use of separators (in the case of English and French, the spaces and punctuation). We may therefore define a segmentation according to predefinite rules : 猫 ‖ が ‖ 鼠 ‖ を ‖ 食べて ‖ いる ‖。. Or, we could accept : 猫 ‖ が ‖ 鼠 ‖ を ‖ 食べている ‖。. Or, we could opt for a chunk/bunsetsu segmentation : 猫が ‖ 鼠を ‖ 食べている ‖。. Three observations can be made from the last examples : there exists not one but several seemingly "correct" boundaries for words in non-segmenting languages ; other units (such as the chunk/bunsetsu demonstrated here in Japanese) may appear to be as relevant as the word, if more ; a common practice in NLP is to transpose approaches which seem straightforward in segmenting languages, to non-segmenting languages where such approaches may be inappropriate and even counter-intuitive.

The conclusion is that boundaries, and therefore a unit of word may not be precisely defined in the general case, a fact which is in agreement with general linguistic theory. Whereas such a unit may be investigated in the scope of one given language, even in that case the application of rigorous criteria yields an analysis that is often distant from the common use of the term. Indeed, the same issue is present in segmenting languages. For instance, as (Martinet, 1970) points out :

- the case of the English genitive such as in *The King of England's* makes it unclear how many words we count (is it 4 or 5 words?) ;
- in the French compound *Bonne d'enfant* we may count 1 or 3 words, and in its German equivalent *Kindermädchen* we could legitimately count 1, but also 2 words.

The object of this study is therefore to investigate whether the character as a processing unit can yield acceptable results and dispense with segmenting into words in Japanese as well as in English. This proposal is tested hereafter on a chain of NLP applications

3 The character as a unit of processing language

To bypass the issues exposed above, we suggest the use of the character unit, which is immediately accessible in all languages in their written electronic form. Depending on the codeset used, the size of a character may vary from one to several bytes of data. This size being known in a given codeset, we argue that language in its written electronic form should be processed as strings of characters rather than strings of words. Finding separators or boundaries, such as punctuation or the space in Roman languages, therefore becomes unnecessary as punctuation marks and spaces are taken for what they are : mere characters part of a larger string. Another advantage of such unit is that, the character size in a given codeset being accounted for, methods investigated become language independent at least theoretically. Lastly, while longer strings of characters are needed to encompass what would otherwise be a sequence of words, this is compensated by the fact that the vocabulary size becomes smaller. Indeed, in a token experiment on a corpus of $20,000$ lines of English data, a vocabulary of $8,191$ words sizes down to a vocabulary of only 61 characters (with an average of 3.91 characters per word). Similarly, on a corpus of $20,000$ lines of Japanese data run through a segmenter, a vocabulary of $9,506$ words sizes down to a vocabulary of $1,871$ characters (with an average of 1.87 characters per word).

4 Previous works

The problem of finding word boundaries in Chinese is an active area of study (Sproat and Emerson, 2003), as is the case in Japanese, for which most of the NLP community relies on the popular tool ChaSen (Matsumoto et al., 2002). The use of the character unit has been investigated in several previous studies : (Dunning, 1994) showed the interest of character based models for language identification in the way that the required training and test sets are surprisingly small in order to achieve good results. (Cercone, 2005) exposes several applications of processing language using the unit of characters, albeit only on English : authorship attribution through common character sequences, analyzing spontaneous speech transcripts in the case of Alzheimer dementia, and the detection of malicious code in data. (Sornlertlamvanich and Tanaka, 1996) investigate ways of extracting words satistically in non-segmenting languages, through the use of left and right mutual information in strings of characters. In the following section we demonstrate the effective use of the character unit on a range of NLP applications.

5 Range of NLP application

5.1 Linguistic data processing : automatic profiling at the character level

Comparing and quantifying corpora is a key issue in corpus based NLP, for which there is a lack of automatic measures. This makes it hard for a user to isolate, transpose, or extend the interesting features of a corpus to other NLP systems. Using the character unit, we addressed the issue of measuring similarity between corpora in (Denoual, 2006). We suggested the setup of a scale between two chosen corpora on which any third given corpus could be assigned a coefficient of similarity, based on the cross-entropy of statistical N-gram character language models.

A possible application of this framework is to quantify similarity in terms of literality, without using word-segmentation. To this end we carried out experiments on several well-known corpora in both English and Japanese language, and showed that the defined similarity coefficient is robust in terms of context length variations, and of the language used.

As an example of application, in both English and Japanese language, we used as references a corpus of oral transcripts at one end, and at the other a corpus of literature and newspapers. We tested our scale on corpora originating from various backgrounds : MAD (Machine Aided Dialogue) is a collection of transcripts of scenarised dialogues, BTEC (Basic Travel Expression Corpus) is a collection of utterances originating from travel phrasebooks, SPAE/NHK are collections of utterances of highly scenarised formal speech, and TIME/Mainichi collections of news texts. Table 1 sums up the results for 5-gram character models. The fast computation of such measure allows in this specific case to separate clearly corpora originating from speech, from corpora originating from the written domain, and quantify their differences.

表 1: Coefficient of literality for 5-gram character models.

	MAD	BTEC	SPAE NHK	TIME Mainichi
Eng	0.22	0.32	0.47	0.51
Jpn	0.38	0.34	0.49	0.49

The approach relying on the character level shows similar performance to other existing similarity measures that use the word unit, in English as well as in Japanese. The unit of character therefore allows faster and more economical processing of linguistic data without the need of language specific word-segmentation.

5.2 Machine translation : character level analogy-based EBMT

We designed, implemented and assessed ALEPH in (Lepage and Denoual, 2005b) an EBMT system that uses a specific operation : proportional analogy. The operation exploits equations of the form

$$A : B :: C : x$$

where A, B, and C are character strings. The equation may then yield a solution $x = D$ in the form of another character string.

This operation implicitly neutralises divergences between languages and captures lexical and syntactical variations along the paradigmatic and syntagmatic axes, without explicitly decomposing sentences into fragments. The system does not require any preprocessing of the aligned examples, such as a step of segmentation into words, as it operates strictly at the character level. While translating, proportional analogy distributes the information at a lower level than that of words, all over the target string of characters.

We conducted an experiment of Japanese to English translation with a test set of 510 input sentences using an unsegmented corpus of almost $160,000$ aligned sentences in Japanese and English. Table 2 shows the scores obtained along with a Gold Standard (reference translation run through the evaluation metric) and a baseline consisting of a translation memory based on edit distance.

表 2: Scores for the Gold Standard, the system and the baseline (translation memory).

System	BLEU	NIST	mWER	GTM
Gold Std	1.00	14.946	0.00	0.915
System	0.600	8.934	0.352	0.721
Baseline	0.378	7.538	0.582	0.611

Such results are on par with state of the art MT engines, and demonstrate the successful implementation of an EBMT system performing all computation at the character level with no need of prior word-segmentation of training data, either on the source side (Japanese) or on the target side (English).

5.3 MT evaluation : an objective measure at the character level

Automatic evaluation metrics for Machine Translation (MT) systems, such as BLEU (Papineni et al., 2001) or NIST (Doddington, 2002), are now well established. They serve as quality assessment methods or comparison tools and are a fast way of measuring improvement. Although it is claimed that such objective MT evaluation methods are language-independent, they are usually only applied to English, as they basically rely on word counts. The organisers of campaigns such as NIST (Przybocki, 2004)[1], TIDES[2] or IWSLT (Akiba et al.,)[3], prefer to evaluate outputs of machine translation systems which are already segmented into words before applying objective evaluation methods. As a consequence, evaluation campaigns of English to Japanese or English to Chinese machine translation systems for instance, are not widely seen or reported.

In (Denoual and Lepage, 2005) we established the equivalence between the standard use of BLEU in word n-grams and its application at the character level in character m-grams. The use of BLEU in character units eliminates the word-segmentation problem : it makes it possible to directly compare commercial systems outputting unsegmented texts with for instance, statistical MT systems which usually segment their outputs. Our study on the English language showed a high correlation, a good agreement in judgement, and an analogy of behaviour for definite corresponding values of n and m. For the most widely used value of n, 4, we determined a corresponding value in characters of $m = 18$. On an ongoing study on the Japanese language, for the most widely used value of n, 4, we determined a corresponding value in characters of $m = 9$.

This paves the way to the application of BLEU in character units in objective evaluation campaigns of machine translation into languages without word delimiters, like Chinese or Japanese.

5.4 Automatic paraphrasing : character based generation

In (Lepage and Denoual, 2005a) we presented and evaluated a method to automatically produce paraphrases from seed sentences, from a given linguistic resource. Our aim is to generate paraphrases to serve as reference sets of MT evaluation measures such as BLEU and NIST.

Lexical and syntactical variation among paraphrases is handled through commutations exhibited in the proportional analogies relying only on character strings (as explained above in the Machine translation section). Well-formedness is enforced by filtering with sequences of characters, not words, of a certain length.

In an experiment, the quality of the paraphrases produced, *i.e.*, (i) their grammaticality, (ii) their equivalence in meaning with the seed sentence, and,

[1] http://www.nist.gov/speech/tests/mt/doc-/mt04_evalplan.v2.1.pdf

[2] http://www.nist.gov/speech/tests/mt-/mt_tides01_knight.pdf

[3] http://www.slt.atr.jp/IWSLT2004/archives-/000619.html

(iii) the internal lexical and syntactical variation in a set of paraphrases, was assessed by sampling and objective measures. With a linguistic resource of 97,769 sentences we generated 8.65 paraphrases in average for 16,153 seed sentences. The grammaticality was evaluated by sampling and was shown to be of at least 99% grammatically and semantically correct sentences (p-value = 2.22%), a quality comparable to that of the original linguistic resource. In addition, at least 96% of the candidates (p-value = 1.92%) were correct paraphrases either by meaning equivalence or entailment.

The lexical and syntactical variation was assessed using BLEU and NIST, and showed slightly better variation than reference sets produced by hand. The figures imply that such generation of paraphrases may become less costly than generating translation references by hand, and with better results.

6 Conclusion

We have shown in this study the interest of using the character unit in natural language processing, instead of a hypothetical word unit which may not be immediately accessible or relevant in the case of non-segmenting languages such as Japanese or Chinese. The possible use of such a unit was demonstrated on a chain of NLP applications : linguistic data profiling, machine translation and its evaluation, and automatic paraphrase generation, to good results. This study argues that the detection of word boundaries and more generally the systematic segmentation of written language into units of words may not be justified, either theoretically or in regard to actual system performance.

7 Acknowledgements

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled “A study of speech dialogue translation technology based on a large corpus”.

参考文献

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. Overview of the IWSLT04 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Nick Cercone. 2005. From simple techniques to impressive results. In *Proceedings of PACLING*, pages 58–66, Tokyo, Japan.

Etienne Denoual and Yves Lepage. 2005. Bleu in characters: towards automatic mt evaluation in languages without word delimiters. In *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing*, pages 81–86, Jeju, Korea.

Etienne Denoual. 2006. A method to quantify corpus similarity and its application to quantifying the degree of literality in a document. *International Journal of Technology and Human Interaction*, 2(1):51–66, January.

George Doddington. 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of Human Language Technology*, pages 128–132, San Diego, March.

Ted Dunning. 1994. Statistical identification of language. Rapport technique, New Mexico State University.

Yves Lepage and Etienne Denoual. 2005a. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05) International Workshop on Paraphrasing (IWP 2005)*, pages 57–64, Jeju, Korea.

Yves Lepage and Etienne Denoual. 2005b. The 'purest' ever built ebmt system: no variable, no template, no training, examples, just examples, only examples. In *Proceedings of the MT Summit X Workshop on Example-Based Machine Translation*, pages 81–80, Phuket, Thailand.

André Martinet. 1970. *Éléments de linguistique générale*. coll. Cursus. Armand Colin, Paris.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. 2002. Morphological analysis system chasen version 2.2.9. Manuel, Nara Institute of Science and Technology.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Research report RC22176, IBM, September.

Mark Przybocki. 2004. The 2004 NIST machine translation evaluation plan (MT-04). Plan, NIST, July.

Virach Sornlertlamvanich and Hozumi Tanaka. 1996. The automatic extraction of open compounds from text corpora. In *Proceedings of the 16th conference on Computational linguistics*, pages 1143–1146, Morristown, NJ, USA. Association for Computational Linguistics.

Richard Sproat and Thomas Emerson. 2003. The first international chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japon.