INTEGRATING WORD BOUNDARY IDENTIFICATION WITH SENTENCE UNDERSTANDING

Kok Wee Gan
Department of Information Systems & Computer Science
National University of Singapore
Kent Ridge Crescent, Singapore 0511
Internet: gankw@iscs.nus.sg

Abstract

Chinese sentences are written with no special delimiters such as space to indicate word boundaries. Existing Chinese NLP systems therefore employ preprocessors to segment sentences into words. Contrary to the conventional wisdom of separating this issue from the task of sentence understanding, we propose an integrated model that performs word boundary identification in lockstep with sentence understanding. In this approach, there is no distinction between rules for word boundary identification and rules for sentence understanding. These two functions are combined. Word boundary ambiguities are detected, especially the fallacious ones, when they block the primary task of discovering the inter-relationships among the various constituents of a sentence, which essentially is the essence of the understanding process. In this approach, statistical information is also incorporated, providing the system a quick and fairly reliable starting ground to carry out the primary task of relationship- building.

1 THE PROBLEM

Chinese sentences are written with no special delimiters such as space to indicate word boundaries. Existing Chinese NLP systems therefore employ preprocessors to segment sentences into words. Many techniques have been developed for this task, from simple pattern matching methods (e.g., maximum matching, reverse maximum matching) (Wang, et al., 1990; Kang & Zheng, 1991), to statistical methods (e.g., word association, relaxation) (Sproat & Shih, 1990; Fan & Tsai, 1988), to rule-based approaches (Huang, 1989 ; Yeh & Lee, 1991; He, et al., 1991).

However, it is observed that simple pattern matching methods and stochastic methods perform poorly in sentences such as (1), (2), and (3), where word boundary ambiguities exist. [1]

(1)	ta	*benren*	*sheng*	le
	She	alone	give birth to	ASP
	san	ge	haizi	
	three	CL	child	

She alone gives birth to three children.

(2)	ta	zhi	kao	dao	*shi*	*fen*
	H/She	only	score	up to	ten	mark

H/She scores only ten marks.

(3)	zhongguo	yi	kaifa	*he*
	China	already	develop	and
	shang	wei	kaifa	de
	yet	not	develop	ASSOC
	shiyou	ziyuan	hen	duo
	oil	resource	very	many

There are many developed and not yet developed oil resources in China.

This problem can be dealt with in a more systematic and effective way if syntactic and semantic analyses are also incorporated. The frequency in which this problem occurs justifies the additional effort needed. However, contemporary approaches of constructing a standalone, rule-based word segmentor do not offer the solution, as this would mean duplicating the effort of syntactic and semantic analyses twice: first in the preprocessing phase, and later in the understanding phase. Moreover, separating the issue of word boundary identification from sentence understanding often leads to devising word segmentation rules which are arbitrary and word specific, [2] and hence not useful at all for sentence understanding. Most importantly, the rules devised always face the problem of over-generalization.

Contrary to conventional wisdom, we do not view the task of word boundary identification as separated from the task of sentence understanding. Rather, the former is regarded as one of the tasks an NLP system must handle within the understanding phase. This perspective allows us to devise a more systematic and natural solution to the problem, at the same time avoiding the duplication of morphological, syntactic, and semantic analyses in two separate stages: the preprocessing stage and the understanding stage.

The basic principle underlying this approach is: every constituent in a sentence must be meaningfully related (syntactically and/or semantically) to some other constituent. Understanding a sentence is simply a process to discover this network of relations. A violation of this principle signifies the presence of abnormal groupings (fallacious word boundaries), which must be removed. [3] For example, the fallacious grouping *rensheng* 'life', if it exists in (1), can be detected by observing a violation of the syntactic relation between this group and *le*, which is

[1] The ambiguous fragments in italics in (1), (2), and (3), *benren sheng*, *shi fen*, and *he shang*, will be wrongly identified as: *ben rensheng*, *shifen*, and *heshang*, respectively, by statistical approaches.

[2] For example, a heuristic rule to resolve the ambiguous fragment *shi fen* in (2): adverb *shifen* 'very' cannot occur at the end of a sentence. This rule rules out the grouping *shifen* to appear in sentence (2).

[3] This principle, in its present form, is too tight for handling metonymic usage of language, as well as ill-formed sentences. We will leave this for future work.

an aspect marker that cannot be a nominal modifier. In (2), selectional restrictions on the RANGE of the verb *kao*, which must either be pedagogical (e.g., *kao shuxue* 'test Mathematics'), resultative (e.g., *kao shibai le* 'test fail ASPECT'), or time (e.g., *kao le yi ge xingqi* 'test ASPECT one week'), rules out the grouping *shifen* 'very', which is a degree marker. [4] Sentence (3) also requires thematic role interpretation to resolve the ambiguous fragment. Selectional restrictions on the PATIENT of the verb *kaifa* 'develop', which must be either a concrete material (e.g., *kaifa meikuang* 'develop coal mine') or a location (e.g., *kaifa sanqu* 'develop rural area'), rules out interpreting the ambiguous fragment *he shang* as *heshang* 'monk'. [5]

This approach, however, does not totally discard the use of statistical information. On the contrary, we use statistical information [6] to give our system a quick and fairly reliable initial guess of the likely word boundaries in a sentence. Based on these suggested word boundaries, the system proceeds to the primary task of determining the syntactic and semantic relations that may exist in the sentence (i.e., the understanding process). Any violation encountered in this process signals the presence of abnormal groupings, which must be removed.

Our approach will not lead to an exceedingly complex system, mainly because we have made use of statistical information to provide us the initial guide. It does not generate all possible word boundary combinations in order to select the best one. Rather, alternative paths are explored only when the current one leads to some violation. This feature makes its complexity not more than that of a two-stage system where syntax and semantics at the later stage of processing signal to the preprocessor that certain lexemes have been wrongly identified.

2 THE PROPOSED MODEL

The approach we proposed takes in as input a stream of characters of a sentence rather than a collection of correctly pre-segmented words. It performs word boundary disambiguation concurrently with sentence understanding. In our investigation, we focus on sentences with clearly ambiguous word boundaries as they constitute an appropriate testbed for us to investigate the deeply interwoven relationships between these two tasks.

Since we are proposing an integrated approach to word boundary identification and sentence understanding, conventional sequential-based architectures are not appropriate. A suitable computational model should have at least the following features: (i) linguistic information such as morphology, syntax, and semantics should be available simultaneously so that it can be drawn upon whenever necessary; (ii) the architecture should allow competing interpretations to coexist and give each one a chance to develop; (iii) partial solutions should be flexible enough that they can be easily modified and regrouped; (iv) the architecture can support localized inferencing which will eventually evolve into a global, coherent interpretation of a sentence.

We are using the Copycat model (Hofstadter, 1984; Mitchell, 1990), which has been developed and tested in the domain of analogy-making. There are four components in this architecture: the conceptual network (encodes linguistic concepts), the workspace (the working area), the coderack (a pool of *codelets* waiting to run), and the temperature (controls the rate of understanding). Our model will differ from NLP systems with a similar approach (Goldman, 1990; Hirst, 1988; Small, 1980) primarily through the incorporation of statistical methods, and the nondeterministic control mechanism used. [7] For a detailed discussion, see (Gan, et al., 1992). In essence, this model simulates the understanding process as a crystallization process, in which high-level linguistic structures (e.g., words; analogous to crystals) are formed and hooked up in a proper way as characters (ions) of a sentence are gradually cooled down.

3 AN EXAMPLE

We will use sentence (1) to briefly outline how the model works. [8]

(1) ta benren sheng le san ge haizi [9]

- bottom-up structure building

  The system starts with bottom-up, character-based codelets in the coderack whose task is to evaluate the associative strength between two neighboring characters. [10] One of the codelets will be chosen probabilistically to run. [11] The executing codelet selects an object from the workspace and tries to build some structures on it. For

[4] Notice the difference between this knowledge and the one mentioned in footnote 2. Both are used to disambiguate the fragment *shi fen*. The former is more ad hoc while ours comes in naturally as part and parcel of thematic role interpretation.

[5] We would like to stress that rules in this approach are not distinguished into two separate classes, one for resolving word boundary ambiguities and the other for sentence understanding. Ours combine these two functions together, performing word boundary identification alongside with sentence understanding. We will give a detailed description on the effectiveness of the various kinds of information after we have completed our implementation.

[6] See Section 3 for an example.

[7] See also footnote 11.

[8] Our description here is oversimplified. Many important issues, such as the representation of linguistic knowledge, the treatment of ambiguous fragments that have multiple equally plausible word boundaries, are omitted. The example discussed in this section is a hand-worked test case which is currently being implemented.

[9] The English glosses and translation are omitted here, as they have been shown in Section 1.

[10] The association between two characters is measured based on mutual information (Fano, 1961). It is derived from the frequency that the two characters occur together versus the frequency that they are independent. Here, we find that statistical techniques can be nicely incorporated into the model. We will derive this information from a corpus of 46,520 words of total usage frequency of 13019,814 given to us by Liang Nanyuan of the Beijing University of Aeronautics and Astronautics.

[11] This is another way statistics is used. The selection of which codelet to run, and the selection of which object to work on are decided probabilistically depending on the system temperature. This is the nondeterministic control mechanism mentioned in Section 2.

example, it may select the last two characters *hai* and *zi* in (1) and evaluate their associative strength as equal to 13.34. This association is so strong that another codelet will be called upon to group these two characters into a word-structure, which forms the word *haizi* 'children'.

- top-down influences

  The formation of the word-structure *haizi* activates the WORD [12] node in the network of linguistic concepts. This network is a dynamic controller to ensure that bottom-up processes do not proceed independently of the system's understanding of the global situation. The activation of the WORD node in turn causes the posting of top-down codelets scouting for other would-be word-structures. Thus, single-character words such as *ta* 'she', *le* (aspect marker), *san* 'three', and *ge* (a classifier) may be discovered.

- radical restructuring

  The characters *ren* and *sheng* will be grouped as a word *rensheng* 'life' by bottom-up, character-based codelets, as the associative strength between them is strong (3.75). This is incorrect in (1). It will be detected when an ASPECT-relation builder, spawned after identifying *le* as an aspect marker, tries to construct a syntactic relation between the word-structure *rensheng* 'life' and the word-structure *le* (ASPECT). Since this relation can only be established with a verb, a violation occurs, which causes the temperature to be set to its maximal value. The problematic structure *rensheng* will be dissolved, and the system proceeds in its search for an alternative, recording down in its memory that this structure *rensheng* should not be tried again in future. [13]

4 SUMMARY

In this model, there is an implicit order in which codelets are executed. At the initial stage, the system is more concerned with identifying words. After some word-structures have been built, other types of codelets begin to decipher the syntactic and semantic relations between these structures. From then on, the word identification and higher-level analyses proceed hand-in-hand. In short, the main ideas in our model are: (i) a parallel architecture in which hierarchical, linguistic structures are built up in a piecemeal fashion by competing and cooperating chains of simple, independently acting codelets; (ii) a notion of fluid re-conformability of structures built up by the system; (iii) a parallel terraced scan (Hofstadter, 1984) of possible courses of action; (iv) a temperature variable that dynamically adjusts the amount of randomness in response to how happy the system is with its currently built structures.

ACKNOWLEDGMENTS

This paper will not be in its present form without the invaluable input from Dr. Martha Palmer. I would like to express my greatest thanks to her. I would also like to thank Guojin, Wu Jianhua, Paul Wu, and Wu Zhibiao for their feedback on an earlier draft.

REFERENCES

Fan, C. K. and Tsai, W. H. (1988) Automatic word identification in Chinese sentences by the relaxation technique. Computer Processing of Chinese and Oriental Languages, 4(1):33-56.

Fano, R. (1961) Transmission of information. MIT Press, Cambridge MA.

Goldman, R. (1990) A probabilistic approach to language understanding. PhD thesis, Department of Computer Science, Brown University.

Gan, K. W., Lua, K. T. and Palmer, M. (1992) Modeling language understanding as a crystallization process: an application to ambiguous Chinese word boundaries identification. Technical Report TR50/92, Department of Information Systems and Computer Science, National University of Singapore.

He, K. K, Xu, H. and Sun, B. (1991) Design principle of expert system for automatic words segmentation in written Chinese. Journal of Chinese Information Processing, 5(2):1-14 (in Chinese).

Hirst, G. (1988) Resolving lexical ambiguity computationally with spreading activation and polaroid words. In S. L. Small, G. W. Cottrell, M. K. Tanenhaus (Eds.), Lexical ambiguity resolution, perspectives from psycholinguistics, neuropsychology and artificial intelligence; Morgan Kaufmann Publishers, San Meteo, California, 73-107.

Hofstadter, D. R. (1984) The Copycat project: an experiment in non-determinism and creative analogies. AI Memo No. 755, Massachusetts Institute of Technology, Cambridge, M. A.

Huang, X. X. (1989) A "produce-test" approach to automatic segmentation of written Chinese. Journal of Chinese Information Processing, 3(4):42-48 (in Chinese).

Kang, L. S. and Zheng, J. H. (1991) An algorithm for word segmentation based on mark. In Proceedings of the 10th anniversary of the Chinese Information Processing Society, Beijing, 222-226 (in Chinese).

Mitchell, M. (1990) COPYCAT: a computer model of high-level perception and conceptual slippage in analogy-making. PhD. Dissertation, University of Michigan.

Small, S. L. (1980) Word expert parsing: a theory of distributed word-based natural language understanding. PhD. dissertation, University of Maryland.

Sproat, R. and Shih, C. L. (1990) A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 4(4):336-351.

Wang, Y. C., Su, H. J. and Mo, Y. (1990) Automatic processing Chinese word. Journal of Chinese Information Processing, 4(4):1-11 (in Chinese).

Yeh, C. L. and Lee, H. J. (1991) Rule-based word identification for Mandarin Chinese sentences - a unification approach. Computer Processing of Chinese and Oriental Languages, 5(2):97-118.

[12] This is a node in the conceptual network, which is activated when the system finds that the word concept is relevant to the task it is currently investigating.

[13] We will skip the implementation details here.