Fluid Annotation: A Granularity-aware Annotation Tool for Chinese Word Fluidity

Shu-Kai Hsieh[1], Yu-Hsiang Tseng[2], Chih-Yao Lee[3], Chiung-Yu Chiang[4]
[1,3,4]Graduate Institute of Linguistics, [2]Department of Psychology
[1,2,3,4]Knowledge-Yielding Language Engineering (KYLE)
National Taiwan University, Taipei, Taiwan
{shukai, seantyh, chiyaolee, sonnejoy}@gmail.com

Abstract

This paper presents a novel word granularity-aware annotation framework for Chinese. Anchored in current functionalist linguistics, this model rearranges the boundary of word segmentation and linguistic annotation, and gears toward a deeper understanding of lexical units and their behavior. The web-based annotation UI also supports flexible annotation tasks for various linguistic and affective phenomena.
Keywords: Wordhood, annotation, Chinese Word Segmentation

1. Introduction

Word segmentation has been one of the most important preprocessing NLP tasks in the pipeline-alike architecture for languages without explicit word delimiter in their written forms. The engineering treatment of word segmentation naturally leads to the requirement of the existing gold standard, including a presumably agreeable standard and a word-segmented corpus based on the standard. Unfortunately, this long-standing rationale does not provide a convincing argument that concurs with current findings of cognitive science.

Words as conventionalized symbols which presents the function by which meaning is attached to form. However, the basic units of cognition are clearly not words. Theoretical and empirical advances in the past decade have revealed that word meanings are only pointers to coherent chunks of encyclopedic knowledge (Malt and Wolff, 2010). In the light of reading task,(Liu et al., 2013) show that Chinese reader did not follow the segmentation rules, and tended to chunk single words into large information units, implying that word meanings sometimes work against the way knowledge is organized in memory.

In this paper, we argue that *word-meaning pair* is fluid in nature, whose granularity (in terms of the length of the word) is influenced by its underlying ontology (paradigmatic dimension), surrounding context (syntagmatic dimension) and real-world application (pragmatic force). Under this view, word segmentation can be considered as *wordhood annotation*, disentangling itself from the *error*-prone role in the NLP pipeline architecture.

2. Review

Word segmentation has been a thorny issue in NLP for many decades. In addition to structural ambiguity resolution and unknown word detection, the current focus is concerned with *propagation error* and *domain adaption*. As the pre-processing task in the pipeline architecture, word segmentation errors can propagate to later processing stages. To handle with this, joint approaches exploiting various machine learning models including the latest neural network have been proposed (Lyu et al., 2016; Shao et al., 2017). Second, it has been recognized that different applications and domains have different calls for different granularities of word segmentation. Recent neural domain adaptation approaches also work through cross-domain embeddings to improve the cross-domain performance (Cai and Zhao, 2016; Zhang et al., 2014). However, a critical examination of the underlying assumption, and their assessment in the light of naturally occurring linguistic data, reveal its inherent contradictions (Taylor, 2012). In the following sections, we introduce the proposed Fluid Annotation model in more details.

3. Fluid Annotation

The scheme of Fluid Annotation comprised of three main components: DeepLexicon, Fluid Segmentation & Tagger, and Annotation UI (Figure 1). Six crucial steps were identified in the scheme: (1) unprocessed text was fed into fluid segmentation and tagging preprocessor, where (2) text was segmented with different granularities, and automatically labeled with possible tags; (3) Annotation UI was provided with these segments and tags, in which (4) annotators could furthermore refine (by regrouping, or dividing) the segmentation with fluid segment tool, or annotate the segmentation (with annotation "brush"), and view the annotation in a natural text context. (5) The annotations created by users were again feed backed to deep lexicon, in which granularities parameters of lexical bundles and update the lexicon tag set table were updated. (6) The updated lexicon would again provided latest information to fluid segmentation & tagging in next session. As a result, a cycle was established where not only the flexibility of linguistic pattern is assured, but annotators' effort cumulated in the process.

3.1. DeepLexicon

DeepLexicon provides all candidate words and tag data associated with the given words used in segmentation and tagging. Distinctively, DeepLexicon featured lemma of different granularity that facilitate fluid segmentation in following steps.

Chinese words has strong tendency to be monosyllabic and disyllabic. However, in segmentation or other practical

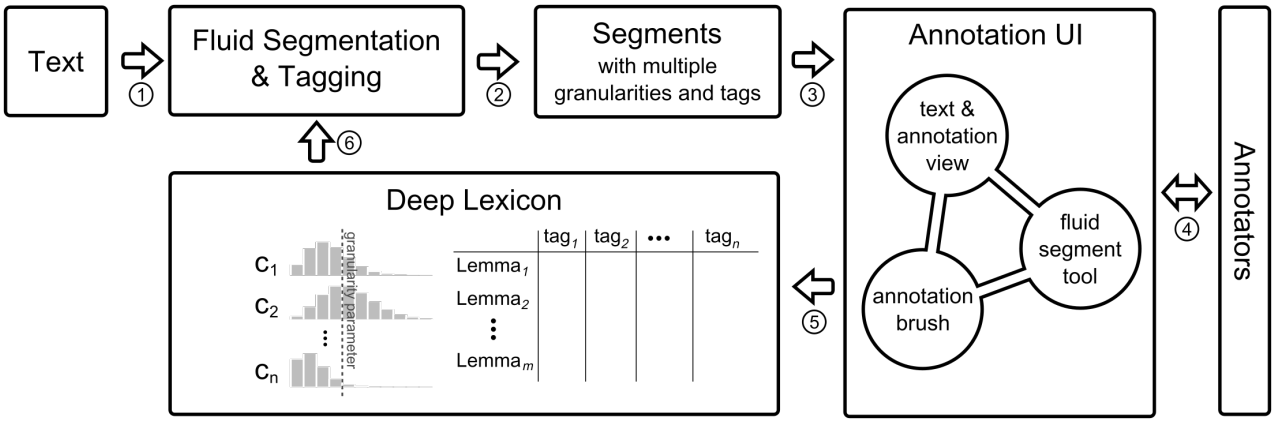

Figure 1: Scheme of Fluid Annotation Overview. Six critical steps were identified in the annotation scheme: (1) input text, (2) preprocess text, (3) prepare segments with multiple granularities and automatic taggings, (4) revise segmentations and tags, (5) annotation feedback, (6) improve segmentation and tagging with new lexical information. Detail descriptions were provided in respective section.

annotation scenarios, word is just one level of information among other linguistic components: multiword expressions, compounds, idiom, or lexical bundles. Previous segmenter relies on a "gold standard" to achieve a high performance in a word segmentation task, virtually eliminate other possibilities to look into groups larger than words.

To alleviate the "hard-cut" issue brought by standard segmentation, DeepLexicon, along with Fulid Segmentation, used in this scheme features "words" of different granularity. "Word granularity" refers to a sequence of lexical patterns of different length. These patterns occurs regularly in different context and carry out a relatively stable communication function. In this sense, "word with different granularities" encompass other linguistic constructs, such as multi-word expression, compounds, idiom, or lexical bundles. For ease of interpretation, granularity is defined as a number ranged from 0 to 1, where we assign granularities of 0 as more fine-grained (shorter patterns, in unit of character count), and 1 as a pattern more coarse-grained (more characters).

In order to operate granularity formally, we further define the granularity of any given word by first calculating the word-length distribution of all the words starting with the same leading character. Secondly, the value of granularity is the cumulative probabilities of the word-length distribution:

$$\text{Granularity}(w) = \sum_{l=1}^{L(w)} p(l; \text{leadChar}(w))$$

where w is the word of interest, leadChar(w) is w's leading character, $L(w)$ denotes the word length of word w and $p(l$;leadChar(w)) is the probability density function of word length l, given the word's leading character.

The current lexicon included 13,5424 lemma which collected from various source. Besides from conventional texts, the lexicon also contained neologism extracted from Taiwan largest Internet forum, emotion expressions and academic lexical bundles commonly found in Chinese academic writings. The resulting word list contained considerably long bundles as revealed in Figure 2. The distribution distinctively called for a novel segmentation procedure which could accommodate the dynamic patterns frequently observed in Chinese discourse.

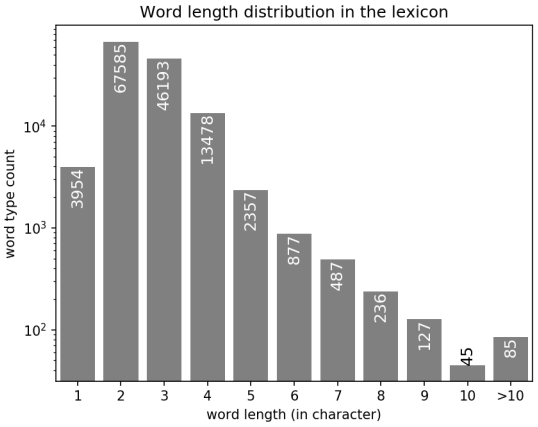

Figure 2: Word length distribution in the lexicon

It is noteworthy that, although the base lexicon already had abundant lexical entries, the lexicon here is designed to be incremental with annotators' collaboration. When annotators group/divide sequence of words in Annotator UI (see below), granularity of the corresponding lemma will automatically adjust accordingly, and segmentation results also reflect the change. Furthermore, we posed few limitations of what can be considered as a "word" in the lexicon. Annotators add their new lemma appropriate in their studies, as long as the pattern is a valid character sequence representable with Unicode. The flexibility is particularly vital when dealing with unconventional and unstructured text, which is dominant in social media, micro-blogging or forums.

Besides the "word" information itself, DeepLexicon also

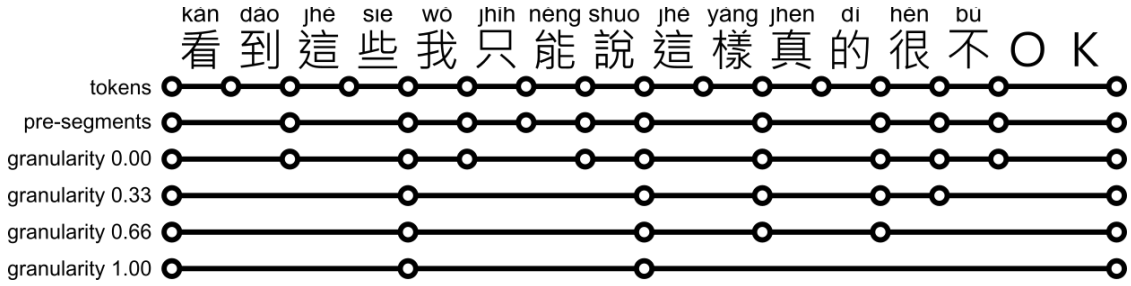

Figure 3: Word segmentations under different granularities.

stored linguistic information from other linguistic resource and user feedback from Annotation UI. For instance, sentiment polarity, mood, and frequency are predefined in DeepLexicon. As annotators created new annotations, these information fed back to DeepLexicon and new tag sets were created. These new tag information along with the predefined tag data, in turn provided a more probable tag by Fluid Tagger. That is, DeepLexicon expanded new lemma and tag information as annotation process progressed. Different annotators could work on the text and share their annotations with others, so the annotations effort could be cumulated in a systematic fashion.

3.2. Segmentation & Tagging

Segmentation is the utmost important step of processing Chinese text. Once the character string of Chinese text has been segmented to multiple words, these words became the only relevant units in subsequent processing steps. The fact that most preprocessing steps only produced single version of word is not without challenged in Chinese, and it profoundly constrained how Chinese text can be annotated and interpreted in later processing steps and analysis.

Fluid Segmenter, instead of pursuing the unique "golden answer", aimed to present the whole spectrum of possibilities on how multiple syllables in Chinese, which represented by multiple individual characters conglomerate into a larger linguistic pattern. Most of Chinese word segmenters based on algorithms which can identify words in a pre-defined segmented corpus. Different segmenters differs on the particular algorithms they implemented. For instance, segmenter in Stanford CoreNLP (Manning et al., 2014) implemented a sophisticated conditional random field model that performed well on segmentation task. However, since the segmenter solely focused on aligning themselves with a predefined word segmentation, different possibilities of segmentation became difficult, if not impossible, to shown themselves in the model outputs.

The segmenter provides words with different granularity by multiple passes of maximal matching and segmentation alignments. To start segmentation, segmenter firstly tries to start with a coarse-grained level (e.g. granularity parameter = 0). Lexicon are queried with the character segmenter encounters, with the granularity parameter in question. The lexicon then offered a full list of possible words starting with the character, whose word granularities are among the designated parameter and 1. The words lexicon provided are then matched against the text from coarse- to fine-grained. If segmenter found a matched, further candidates in the word list are skipped. The segmenter stored the matched sequence, and move the position to the character after the matched sequence. When all characters in the text are attempted, segmenter moved to a higher granularity and repeat the procedures above (Figure 3).

Different granularities of words are identified after the segmenter finish procedures above. These words may contains conflicting word boundaries and isolated single characters which are either one-character words or out-of-vocabularies in lexicon. Although conflicting word boundaries and novel words may itself be an interesting issue in certain research, some studies need an acceptable segmentation so researchers can focus on the patterns of interest.

A pre-segmenter can optionally be incorporated to provide a quick and conventional way of segmentation. The benefits of a pre-segmenter is to solve word ambiguities frequently observed in Chinese text, and alleviate the problem of OOV issues faced in a lexicon-based segmentation. The results from pre-segmenter are the segmentation to which other word granularities align themselves. Specifically, the patterns from different granularities can merge word sequences produced in pre-segment, but dismissed if the word conflict with the pre-segment results. Results from pre-segmenter can be safely ignored, and the final segmentation would only aligned with character-based tokenization.

3.3. Annotation UI

Annotation is a paramount step to develop linguistic theory. Despite the significance in linguistic researches, problems as basic as tokenization still profoundly affect the annotation practices (Ide, 2017). Annotation UI, a browser-based annotation user interface, was aimed to create an environment where researchers could smoothly annotate the focused linguistic phenomena based on the automatic outputs from Fluid Segmentation and Tagger.

Given the fluidity of Chinese expressions, it's unlikely any finite collection of lemma, such as DeepLexicon, could exhaustively satisfied every need of linguistic investigations. Although Fluid Segmentation allow considerably more flexibility to researchers as they can freely decide the level of interest in granularities, there were still some circumstances that annotators or researchers wish finer- or coarser- grained segmentation results, and some of these results are stable across context. These instances are candidates for addition into Lexicon.

Segmentation is one of the most crucial form of anno-

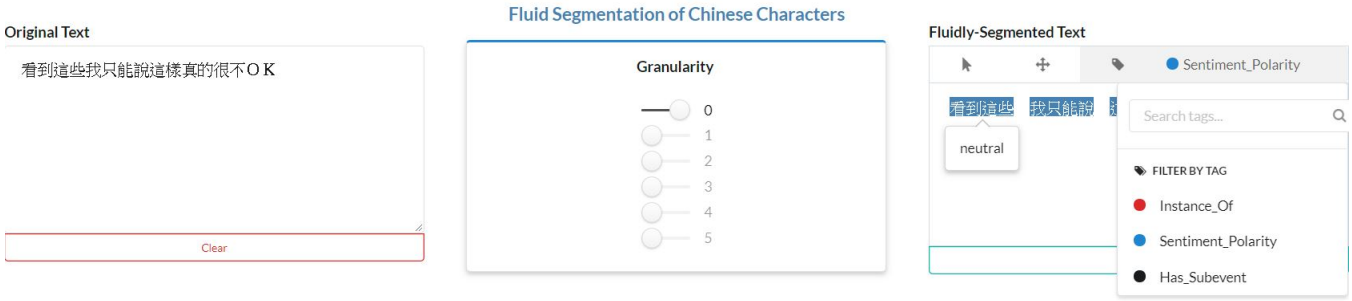

Figure 4: Screenshot of Annotation UI. Left panel was text input window, central panel was for granularity settings, and right panel was annotation window. Three tool buttons were available on the top of the right panel: normal selection tool, fluid segment tool, and annotation brush.

tations. Since most of the subsequent annotations depends on the segmentation in Chinese, the importance of the flexibility supported by segmentation cannot be overstated. Annotators contribute new segments to DeepLexicon through Annotation UI. Annotators first chose an appropriate granularity level. which includes 6 settings, from more coarse-grained (contains longer sequences) to finer-grained (contains shorter sequences), the pre-segment level and the token level. Upon granularity selection, annotators then freely regroup character sequence in text. The annotation process completed when annotators submit the final regrouped text, where lexicon scanned through the segmentation in the text for new patterns. New patterns are added to lexicon and automatically update granularity calculation. These new patterns would be utilized in following text segmentation.

In addition, segments in Annotator UI came with tag suggestion predicted by Fluid Taggers. These suggestions were currently produced by maximum-likelihood estimates based on tag statistics recorded in lexicon, new prediction algorithm can incorporated if more sophisticated suggestion scenario is required. Annotators could either accept the suggested tags, revise the tag, or devised a new tag set entirely. Annotation brush was designed to help annotators intuitively ”paint” the tag on the segments, through which annotators can select categories and tag values they wish to annotate, and click the segments to annotate. Regardless of annotator's decision to add, modify, or delete the tags, annotations would be processed by Annotation UI and feed backed to DeepLexicon, where the tag data would further processed to update future predictions. The procedure ensures the linguistic insights imparted by annotators accumulated in the process.

4. Conclusion

Word segmentation with its underlying generative model of linguistics, and its preprocessing role have set the research agenda for at least half a century in Chinese NLP. Even for languages with space as the word boundary delimiter in their writing system, though useful as a rule of thumb, still begs the question of how words might be defined (Taylor, 2012). Mounting evidence in recent studies have offered alternatively how knowledge comes packaged into coherent chunks in mind, and *word* meaning are closely aligned with these chunks. In the same vein, what we propose in this paper, is a novel fluid annotation model that allows the word granularity to be presented from holistic (un-wordlike) to discrete elements (word-like) via the collaborative annotation. We believe the proposed model could liberate and expand our research imagination, and provides a pathway to connect NLP/NLU with cognitive computing.

5. Bibliographical References

Cai, D. and Zhao, H. (2016). Neural word segmentation learning for chinese. *arXiv preprint arXiv:1606.04300*.

Ide, N., (2017). *Introduction: The Handbook of Linguistic Annotation*, pages 1–18.

Liu, P.-P., Li, W.-J., Lin, N., and Li, X.-S. (2013). Do chinese readers follow the national standard rules for word segmentation during reading? *PloS one*, 8(2):e55440.

Lyu, C., Zhang, Y., and Ji, D. (2016). Joint word segmentation, pos-tagging and syntactic chunking. In *AAAI*, pages 3007–3014.

Malt, B. and Wolff, P. M. (2010). *Words and the mind: How words capture human experience*. Oxford University Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Shao, Y., Hardmeier, C., Tiedemann, J., and Nivre, J. (2017). Character-based joint segmentation and pos tagging for chinese using bidirectional rnn-crf. *arXiv preprint arXiv:1704.01314*.

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford University Press.

Zhang, M., Zhang, Y., Che, W., and Liu, T. (2014). Type-supervised domain adaptation for joint segmentation and pos-tagging. In *EACL*, pages 588–597.