NTU-MC Toolkit: Annotating a Linguistically Diverse Corpus

Liling Tan
Universität des Saarland
Campus, 66123 Saarbrücken, Germany
alvations@gmail.com

Francis Bond
Nanyang Technological University
14 Nanyang Drive, Singapore 637332
bond@ieee.org

Abstract

The NTU-MC Toolkit is a compilation of tools to annotate the Nanyang Technological University - Multilingual Corpus (NTU-MC). The NTU-MC is a parallel corpora of linguistically diverse languages (Arabic, English, Indonesian, Japanese, Korean, Mandarin Chinese, Thai and Vietnamese). The NTU-MC thrives on the mantra of "*more data is better data and more annotation is better information*". Other than increasing parallel data from diverse language pairs, annotating the corpus with various layers of information allows corpora linguists to discover linguistic phenomena and provides computational linguists with pre-annotated features for various NLP tasks. In addition to the agglomeration existing tools into a single python wrapper library, we have implemented three tools (`Mini-segmenter`, `GaChalign` and `Indotag`) that (i) provides users with varying analysis of the corpus, (ii) improves the state-of-art performance and (iii) reimplements a previously unavailable annotation tool as a free and open tool. This paper briefly describes the wrapper classes available in the toolkit and introduces and demonstrates the usage of the `Mini-segmenter`, `GaChalign` and `Indotag`.

1 Introduction

The NTU-MC Toolkit was developed in conjunction with the compilation of the Nanyang Technological University - Multilingual Corpus (NTU-MC) (Tan and Bond, 2012). It is an agglomeration of existing state-of-art tools into a single python wrapper library. The NTU-MC Toolkit provides python wrapper classes for tokenizers and Part-of-Speech (POS) taggers for the respectively languages:

- Stanford Segmenter and POS taggers (Arabic and Chinese)
- POSTECH `POSTAG/K` tagger (Korean)
- `tinysegmenter` and `MeCab` (Japanese)
- `JVnTextPro` (Vietnamese)

Additionally, we implemented three tools to provide complementary or better annotations, viz.:

- `Mini-segmenter` (Chinese): Dictionary based Chinese segmenter
- `GaChalign` (Crosslingual): Gale-Church Sentence-level Aligner with variable parameters
- `Indotag` (Indonesian): Conditional Random Field (CRF) POS tagger.

The following sections of the paper will briefly describe the wrapper classes available in the toolkit (Section 2) and introduce and demonstrate the usage of the `Mini-segmenter` (Section 3), `GaChalign` (Section 4) and `Indotag` (Section 5).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 86–89, Dublin, Ireland, August 23-29 2014.

2 Tokenization and POS Tagger Wrappers

Python wrapper classes were written for (i) Stanford Segmenter and POS tagger (Chang et al., 2008; Toutanova et al., 2003), (ii) POSTECH `POSTAG/K` tagger (Lee et al., 2002), (iii) `tinysegmenter` and `MeCab` (Kudo et al., 2004) and (iv) `JVnTextPro` (Nguyen et al., 2010). Although scientifically uninteresting, it simplifies usage of annotation tools especially for beginner who are new to Natural Language Processing or python programming. The wrapper classes are also compatible with corpora readers of the Natural Language Toolkit (NLTK).

Usage Users can either invoke the wrapper classes programmatically[1]:

```
$ python
»> from ntumc.tk import postech
»> sentence = u"싱가포르에서가장유명한하이난식치킨라이스음식점중하나인Tian
Tian Hainanese Chicken Rice는맥스웰푸드센터(Maxwell Food Centre)에위치해
있으며, 음식점앞에는손님들이매일길게줄지어서있습니다."
»> postech.postagk(sentence)
[(u'싱가포르', 'NNP'), (u'에서', 'JKB'), (u'가장', 'MAG'), (u'유명',
'XR'), (u'하', 'XSA'), (u'ㄴ', 'ETM'), (u'하이난식', 'NNG'), ...]
```

or via command line:

```
$ echo "싱가포르에서가장유명한하이난식치킨라이스음식점중하나인Tian Tian
Hainanese Chicken Rice는맥스웰푸드센터(Maxwell Food Centre)에위치해있으며,
음식점앞에는손님들이매일길게줄지어서있습니다." > input.txt
$ python ntumc/tk/postech.py input.txt > output.txt
```

3 Mini-segmenter

The `mini-segmenter` is dictionary based Chinese segmenter that capitalizes on token length as heuristics for Chinese text tokenization. The tool includes a dictionary of Singaporean Chinese NEs crawled from Wikipedia titles and articles on Singapore.

Motivation The mini-segmenter was created to resolve the problem of segmenting localized Chinese words from the Singaporean variety of Mandarin Chinese in the NTU-MC. After manual inspection, the Stanford Chinese segmenter[2] was segmenting the Chinese tokens with the wrong word boundary. For example, the Stanford Chinese word segmenter wrongly tokenized 乌节路 *wujielu* "Orchard road" as 乌_节路 wu jielu "black joint-road". Originally, these topological terms were re-segmented with a manually crafted dictionary built using Wikipedia's Chinese translations of English names of Singapore places and streets. Then we found more localized Named Entities (NEs) for person names, organizations and food terms. Short of building a manually segmented corpus and retraining the Stanford segmenter models, a simple dictionary approach to segmentation could resolve out-of-domain issue.

Innovation A lightweight lexicon/dictionary based Chinese text segmenter. The advantage of using a lexicon/dictionary for text segmentation is the ability to localize and scale according to the Chinese variety or domain. The `mini-segmenter` ranks the token boundaries based on sum of the square of the tokens' character length, $\sum_i^n len(token_i)^2$, where *n* is the number of tokens and *len(token)* is the character length of each token. This novel scoring is based on the preference for larger chunks than smaller chunks in a sentence.

Usage The full documentation of the `mini-segmenter` can be found on `https://code.google.com/p/mini-segmenter/`

[1]The example sentence in English, "*One of the most famous Hainanese chicken rice stalls in Singapore, Tian Tian Hainanaese Chiken Rice is located in the Maxwell Food Centre, with long queues forming in front of the stall every day.*"

[2]both Penn Chinese Treebank (ctb) and Peking University (pku) models

Results We evaluate the `mini-segmenter` output against the Stanford segmenter output with the `fish-head-curry.txt` from the NTU-MC which was was previously selected at random as a text sample for human annotators to verify the tagger accuracy. The Stanford segmenter with Stanford POS tagger, it achieved 85.94% POS accuracy with 19% mis-segments. Using the `mini-segmenter` with the Stanford POS tagger, it achieved 91.27% POS accuracy with 11.43% mis-segments.

4 GaChalign

The `GaChalign` tool is sentence alignment tool to align sentences given a bitext. The tool is a modification of the original Gale-Church algorithm that capitalized on ratio of characters/tokens of two languages in the bitext to align the sentences (Gale and Church, 1993).

Motivation The Gale-Church algorithm had parameters tuned to suit Indo-European languages more specifically German-English language pairs. When using state-of-art sentence alignment tool based on Gale-Church algorithm to align Chinese, Japanese or Korean texts to their respective English texts, the NTU-MC reported a poor performance in F-measure metrics adheres to standards set by the ARCADE II project (Chiao et al. 2006). We want to see whether it is possible to improve the algorithm by tune algorithm using language-pair specific parameters.

Innovation We replaced the mean, variance and penalty parameters from the Gale-Church algorithm with language-pair specific parameters automatically calculated from a non-aligned corpus.

Results Our experiment with English-Japanese corpus has shown that (i) simply using the calculated character mean from the unaligned text improves precision and recall of the algorithm; from 61.0% (default parameters) to 62.0% (language specific) F-scores) and (ii) using language specific penalties further increased the F-scores to 62.9%. However, aligning syllabic/logographic language (Japanese) to alphabetic language (English) remains a challenge for Gale-Church algorithm[3].

5 Indotag

The `Indotag` is a probabilistic Conditional Random Field (CRF) Bahasa Indonesian Part of Speech (POS) tagger with the specifications recommended by (Pisceldo et al., 2009). The pre-trained model is based on the unigram CRF with 2-left and 2-right context features using the Universitas Indonesia's 1 million word corpus compiled under the Pan Asia Networking Localization (PANL10N) project.

Motivation To reimplement the Indonesian POS tagger described in Pisceldo et al. (2010) using free and open data and licensing it as open source tool.

Innovation None or not much. An open source reimplementation of a Bahasa Indonesian POS tagger.

Result The IndoTag achieved 78% accuracy when annotating the the `fish-head-curry.txt` text sample from the NTU-MC.

6 Discussion

While English POS tagging reports >97% accuracy (Manning, 2011) and sentence alignments for Indo-European languages performs well at >96% (Gale and Church, 1993; Varga et al., 2007), there is much room for improvement with regards to POS tagger accuracy for Asian languages and automatic sentence alignments from syllabic/logographic languages to alphabetic ones. Even though the languages in the NTU-MC are not considered low-resource languages, the tools to annotate them have limited performance. While the maintainers of the NTU-MC continues to push the performance of the individual tools for these languages, we urge researchers to work on improving NLP tools/application for Asian languages.

[3]Detailed evaluation on the `GaChalign` experiments can be found on `https://code.google.com/p/gachalign/`

7 Conclusion

We have introduced the NTU-MC Toolkit that was compiled to annotated the linguistically diverse NTU-MC. The toolkit agglomerate existing tools into a single python wrapper library. The toolkit also implemented the novel dictionary-based segmenter (`Mini-segmenter`) to improve state-of-art performance for Chinese segmentation, an modified Gale-Church algorithm (`GaChalign`) to improve sentence alignments for syllabic-alphabetic language pairs and reimplemented an open source `Indotag` Bahasa Indonesian POS tagger.

Acknowledgements

This research was partially funded by a joint JSPS/NTU grant on Revealing Meaning through Multiple Languages and the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2009-5259-5.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232. Association for Computational Linguistics.

William A. Gale and Kenneth Ward Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *EMNLP*, pages 230–237.

Gary Geunbae Lee, Jeongwon Cha, and Jong-Hyeok Lee. 2002. Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of korean. *Computational Linguistics*, 28(1):53–70.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer.

Cam-Tu Nguyen, Xuan-Hieu Phan, and Thu-Trang Nguyen. 2010. Jvntextpro: A java-based vietnamese text processing tool. `http://jvntextpro.sourceforge.net/`.

Femphy Pisceldo, Ruli Manurung, and Mirna Adriani. 2009. Probabilistic part-of-speech tagging for bahasa indonesia.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). In *International Journal of Asian Language Processing, 22(4)*, page 161–174.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Daniel Varga, Peter Halacsy, AndraS Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor TrOn. 2007. Parallel corpora for medium density languages. *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, 292:247.