A semi-automatic method for constructing MUSE sentiment-annotated corpora

Byoung-Yeol Chae, Dong-Hee Cho, Sairom Kim, Eric Laporte, Jeesun Nam

A semi-automatic method for constructing MUSE sentiment-annotated corpora

Byoung-Yeol Chae*, Dong-hee Cho*, Sairom Kim*, Eric Laporte†, Jeesun Nam‡

* Hankuk University of Foreign Studies
{pfcchae, j.donghee, steviasr}@gmail.com

†LIGM, Université Paris-Est-Marne-la-Vallée
eric.laporte@univ-paris-est.fr

‡Hankuk University of Foreign Studies
namjs@hufs.ac.kr

This study describes a methodology we adopted for constructing Multilingual Sentiment-Annotated Corpora (named MUSE), that consist of two types of annotated corpora: Sentence-based Sentiment-Annotated Corpora (MUSE-SESAC) and Token-based Sentiment-Annotated Corpora (MUSE-TOSAC).

Sentiment-annotated corpora are essential for training and domain adaptation of sentiment analysis systems, as well as for corpus-linguistic studies based on real world data. In this respect, how to create annotations of consistently high quality is an important issue, and with the growth in internet connectivity, using crowdsourcing such as Amazon Mechanical Turks for corpus annotation has become a widespread practice. Nonetheless, according to the language of the texts and the fineness required, crowdsourcing cannot always be the best alternative. In this study, we created sentiment-annotated social-web corpora in Korean by building fine-grained annotation guidelines, training Korean linguistic experts and implementing effective environments for annotation-related tasks.

The methodology proposed for building the MPQA corpus (Wiebe *et al.* 2002, 2005), one of the well-known sentiment-annotated corpora for English, has been adopted for a Korean sentiment-annotated corpus (Shin *et al.* 2012). However, the latter ended in a small-sized annotated corpus (7,713 sentences in total and only 2,658 subjective or sentiment-annotated sentences) and is built on a Korean newspaper corpus, not on user-generated subjective texts. Therefore, in order to satisfy the current needs for annotated corpora, we had to conceive a novel methodology for constructing Korean sentiment-annotated corpora.

In this study, we delimited 4 domains: politics, economy, culture and services, we selected major keywords for each domain and we collected a raw corpus of tweets. About 410 keywords were selected and more than 1 million tweets were collected by these keywords.

The annotation of the corpus is in progress. On one hand, to construct the sentence-based (SESAC) corpus, 6 types of sentiment classification were defined such as {Positive}, {Negative}, {Neutral}, {Contradictory}, {Objective} and {Trash}. On the other hand, to construct the token-based (TOSAC) corpus, 2 phases were planned: DECO-based Semiautomatic Annotation (DESA Phase) and TOOL-based Manual Annotation (TOMA Phase). In the DESA phase, a corpus analysis platform named UNITEX (Paumier 2003) is used to apply a Korean Electronic Sentiment Lexicon, SELEX (http://dicora.hufs.ac.kr), to our raw corpus. The application of SELEX to the corpus will provide an automatic annotation with 7 sentiment classification tags: {QXSP (Strongly Positive)}, {QXPO (Positive)}, {QXNE (Neutral)}, {QXNG (Negative)}, {QXSN (Strongly Negative)}, {QXDE ((Context)-Dependent)} and {QXAD (Accentuated Dependency)} and 9 tags of named entity that may be a target of sentiment expressions: {XXPE (Person)}, {XXOR (Organization)}, {XXGE

(Geography)}, {XXLO (Location)}, {XXTI (Time)}, {XXEV (Event)}, {XXCO (Concrete)}, {XXPR (Product)} and {XXCR (Creation)}.

The TOMA phase is performed with an annotation tool, DecoLex (http://dicora.hufs.ac.kr), which opens the results of the DESA phase and permits the human annotators to modify the annotation and add the Sentiment Annotation Elements predetermined in this study on the basis of Liu's Opinion Quintuple (2012).

In this way, we constructed a SESAC corpus of 165,000 tweets and a TOSAC corpus of 30,000 tweets presented in XML format. The methodology proposed in this study will be applied to building multilingual sentiment-annotated corpora in our future works.

References

Callison-Burch, Chris. 2009. Fast, cheap, and creative: evaluating translation quality using amazon''s mechanical turk. In EMNLP ''09: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–-295, Morristown, NJ, USA. Association for Computational Linguistics.

Fleischman, Michael. 2001. Automated Subcategorization of Named Entities. In *Proceedings of the Conference of the European Chapter of Association for Computational Linguistics.*

Fleischman, Michael & E. Hovy. 2002. Fine Grained Classification of Named Entities. In *Proceedings of the Conference on Computational Linguistics.*

Kaisser, Michael & John Lowe. 2008. Creating a research collection of question answer sentence pairs with amazons mechanical turk. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC''08).*

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers.

Maynard, Diana, V. Tablan, C. Ursu, H. Cunningham & Y. Wilks, 2001. Named Entity Recognition from Diverse Text Types. In *Proceedings of the Recent Advances in Natural Language Processing*.

Nadeau, David & Satoshi Sekine. 2007. A survey of named entity recognition and classification. Named Entities: Recognition, classification and use. *Lingvisticæ Investigationes*, 30:1. John Benjamins Publishing Company. 3–26

Nam, Jeesun.2010. *Korean Electronic Dictionary DECO*. DICORA Technical Report TR-2010-02, Hankuk University of Foreign Studies.

Nam, Jeesun.2015. *Korean Electronic Dictionary DECO*. DICORA Technical Report TR-2015-02, Hankuk University of Foreign Studies.

Pak, Alexander & Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, In *Proceedings of the European Language Resources Association(ELRA).*

Paumier, Sébastien. 2003. *Unitex Users' Manual*. UPEM, France.

Shin, Hyopil, Munhyong Kim, Yumi Jo, Hayeon Jang & Andrew Cattle. 2012. Annotation Scheme for Constructing Sentiment Corpus in Korean. In *Proceedings of the 26th Pacific Asia Conference on Language*, Universitas Indonesia, 181-190.

Sorokin, A. & D. Forsyth. 2008. *Utility data annotation with amazon mechanical turk*.

Wiebe, Janyce. 2002. *Instructions for annotating opinions in newspaper articles*. Department of Computer Science Technical Report TR-02-101, University of Pittsburgh.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* (formerly Computers and the Humanities), 39(2/3):164–210.