The cross-lingual sentiment classification

Science

Cross-lingual sentiment classification means that the systems performing sentiment classification can be applied to more than one language. This approach is really promising.

First, the imbalance of opinionated-rich resources in different languages is pretty drastically because of the available number of corpora and lexicons. Building a highly accurate sentiment classifier in English is often easier and more favorable than creating one in Vietnamese since resources of corpus and lexicons are still poor and far from being comparable to what English can offer. The limitation of resources and tools has arguably slowed down the process of building a reasonably accurate sentiment classifier in Vietnamese. However, with this approach, those resources in English are made available to other languages through machine translation.

Second, when the business domains of a company spreading overseas, there are high chances that they gather data not only in their native language but also from their foreign markets, which are really useful for them to extend their business. However, this approach is questionable in the manner that whether machine translated materials and those mentioned sentiment classification tools in English are helpful to build a reasonably accurate sentiment classifier in other languages or not, especially in Vietnamese.

In recent years, there are some notable studies on this topic. However, none of the published studies are performed on Vietnamese, and we believe that it is worth conducting experiments on this topic because this approach is really promising for poor resources language such as Vietnamese. The following two methods are document-level sentiment classification.

In 2008, Wan proposed a co-training approach in cross-lingual sentiment classification which uses annotated English corpus to classify Chinese reviews without using any Chinese resources. The training set contains labeled English reviews and unlabelled Chinese reviews. These reviews are then translated into separated English and Chinese version of themselves. With these resources, the author then learns two different classifiers using Support Vector Machine (SVM), a machine learning algorithm, in separated views. Finally, those two classifiers are combined into one classifier for both English and Chinese. For the test set, each unlabelled Chinese reviews are translated into an English one before applying the trained model to classify them into positive or negative class.

Another worth mentioning studies in cross-lingual sentiment classification is Bilingual Document Representation Learning (BiDRL) proposed by Zhou in 2016. This approach uses BiDRL model which simultaneously learns both word and document representation in both source and target languages. The authors also use a joint learning algorithm which exploits both monolingual and bilingual constraints. A monolingual constraint is used to obtain both word and document embedding via extended paragraph vector model for both source and target language in a semi-supervised manner, and a bilingual constraint is used to build the consistency in both semantic and sentiment relationship across both languages.

From the original definition of cross-lingual sentiment classification and the two above methods, we can see that the cross-language sentiment classification approaches require a source language and a destination language acquired by a machine translation process to bridge the language

gap. A source language is the one with a richer opinionated resource that we look up and build our desired sentiment classification system upon. A destination language is we want to build the system for. However, due to the domain differences between different datasets, we cannot obtain any English datasets, which has a similar or close domain to our Vietnamese one, to work as a source language. Therefore, our approach does not follow the common pattern of a pair of source and destination languages-in our case, Vietnamese and English are simply treated as two different views.

Furthermore, we have not seen any sentiment classification system in Vietnam that can perform bilingual sentiment classification; therefore, we believe that our research is worth trying because of the experimental values. Generally, in this thesis, we want, first, to build a system which has an ability to perform sentiment classification on both English and Vietnamese and, second, to prove that translated English materials help to improve the overall accuracy for Vietnamese sentiment classification standing in the cross-lingual point of view.