Identifying High-Impact Sub-Structures for Convolution Kernels in Document-level Sentiment Classification

Zhaopeng Tu[†] **Yifan He**[‡§] **Jennifer Foster**[§] **Josef van Genabith**[§] **Qun Liu**[†] **Shouxun Lin**[†]

[†]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology, CAS

[‡]Computer Science Department
New York University

[§]School of Computing
Dublin City University

[†]{tuzhaopeng,liuqun,sxlin}@ict.ac.cn,
[‡]yhe@cs.nyu.edu, [§]{jfoster,josef}@computing.dcu.ie

Abstract

Convolution kernels support the modeling of complex syntactic information in machine-learning tasks. However, such models are highly sensitive to the type and size of syntactic structure used. It is therefore an important challenge to automatically identify high impact sub-structures relevant to a given task. In this paper we present a systematic study investigating (combinations of) sequence and convolution kernels using different types of sub-structures in document-level sentiment classification. We show that minimal sub-structures extracted from constituency and dependency trees guided by a polarity lexicon show 1.45 point absolute improvement in accuracy over a bag-of-words classifier on a widely used sentiment corpus.

1 Introduction

An important subtask in sentiment analysis is sentiment classification. Sentiment classification involves the identification of positive and negative opinions from a text segment at various levels of granularity including *document-level*, *paragraph-level*, *sentence-level* and *phrase-level*. This paper focuses on document-level sentiment classification.

There has been a substantial amount of work on document-level sentiment classification. In early pioneering work, Pang and Lee (2004) use a flat feature vector (e.g., a bag-of-words) to represent the documents. A bag-of-words approach, however, cannot capture important information obtained from structural linguistic analysis of the documents. More recently, there have been several approaches which employ features based on deep linguistic analysis with encouraging results including Joshi and Penstein-Rose (2009) and Liu and Seneff (2009). However, as they select features manually, these methods would require additional labor when ported to other languages and domains.

In this paper, we study and evaluate diverse linguistic structures encoded as convolution kernels for the document-level sentiment classification problem, in order to utilize syntactic structures without defining explicit linguistic rules. While the application of kernel methods could seem intuitive for many tasks, it is non-trivial to apply convolution kernels to document-level sentiment classification: previous work has already shown that categorically using the entire syntactic structure of a single sentence would produce too many features for a convolution kernel (Zhang et al., 2006; Moschitti et al., 2008). We expect the situation to be worse for our task as we work with documents that tend to comprise dozens of sentences.

It is therefore necessary to choose appropriate substructures of a sentence as opposed to using the whole structure in order to effectively use convolution kernels in our task. It has been observed that not every part of a document is equally informative for identifying the polarity of the whole document (Yu and Hatzivassiloglou, 2003; Pang and Lee, 2004; Koppel and Schler, 2005; Ferguson et al., 2009): a film review often uses lengthy objective paragraphs to simply describe the plot. Such objective portions do not contain the author's opinion and are irrelevant with respect to the sentiment classifi-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 338–343,
Jeju, Republic of Korea, 8-14 July 2012.

cation task. Indeed, separating objective sentences from subjective sentences in a document produces encouraging results (Yu and Hatzivassiloglou, 2003; Pang and Lee, 2004; Koppel and Schler, 2005; Ferguson et al., 2009). Our research is inspired by these observations. Unlike in the previous work, however, we focus on syntactic *substructures* (rather than entire paragraphs or sentences) that contain subjective words.

More specifically, we use the terms in the lexicon constructed from (Wilson et al., 2005) as the indicators to identify the substructures for the convolution kernels, and extract different sub-structures according to these indicators for various types of parse trees (Section 3). An empirical evaluation on a widely used sentiment corpus shows an improvement of 1.45 point in accuracy over the baseline resulting from a combination of bag-of-words and high-impact parse features (Section 4).

2 Related Work

Our research builds on previous work in the field of sentiment classification and convolution kernels. For sentiment classification, the design of lexical and syntactic features is an important first step. Several approaches propose feature-based learning algorithms for this problem. Pang and Lee (2004) and Dave et al. (2003) represent a document as a bag-of-words; Matsumoto et al., (2005) extract frequently occurring connected subtrees from dependency parsing; Joshi and Penstein-Rose (2009) use a transformation of dependency relation triples; Liu and Seneff (2009) extract adverb-adjective-noun relations from dependency parser output.

Previous research has convincingly demonstrated a kernel's ability to generate large feature sets, which is useful to quickly model new and not well understood linguistic phenomena in machine learning, and has led to improvements in various NLP tasks, including relation extraction (Bunescu and Mooney, 2005a; Bunescu and Mooney, 2005b; Zhang et al., 2006; Nguyen et al., 2009), question answering (Moschitti and Quarteroni, 2008), semantic role labeling (Moschitti et al., 2008).

Convolution kernels have been used before in sentiment analysis: Wiegand and Klakow (2010) use convolution kernels for opinion holder extraction, Johansson and Moschitti (2010) for opinion expression detection and Agarwal et al. (2011) for sentiment analysis of Twitter data. Wiegand and Klakow (2010) use e.g. noun phrases as possible candidate opinion holders, in our work we extract any minimal syntactic context containing a subjective word. Johansson and Moschitti (2010) and Agarwal et al. (2011) process sentences and tweets respectively. However, as these are considerably shorter than documents, their feature space is less complex, and pruning is not as pertinent.

3 Kernels for Sentiment Classification

3.1 Linguistic Representations

We explore both sequence and convolution kernels to exploit information on surface and syntactic levels. For sequence kernels, we make use of lexical words with some syntactic information in the form of part-of-speech (POS) tags. More specifically, we define three types of sequences:

- SW, a sequence of lexical words, e.g.: *A tragic waste of talent and incredible visual effects*.
- SP, a sequence of POS tags, e.g.: *DT JJ NN IN NN CC JJ JJ NNS*.
- SWP, a sequence of words and POS tags, e.g.: *A/DT tragic/JJ waste/NN of/IN talent/NN and/CC incredible/JJ visual/JJ effects/NNS*.

In addition, we experiment with constituency tree kernels (CON), and dependency tree kernels (D), which capture hierarchical constituency structure and labeled dependency relations between words, respectively. For dependency kernels, we test with word (DW), POS (DP), and combined word-and-POS settings (DWP), and similarly for simple sequence kernels (SW, SP and SWP). We also use a vector kernel (VK) in a bag-of-words baseline. Figure 1 shows the constituent and dependency structure for the above sentence.

3.2 Settings

As kernel-based algorithms inherently explore the whole feature space to weight the features, it is important to choose appropriate substructures to remove unnecessary features as much as possible.

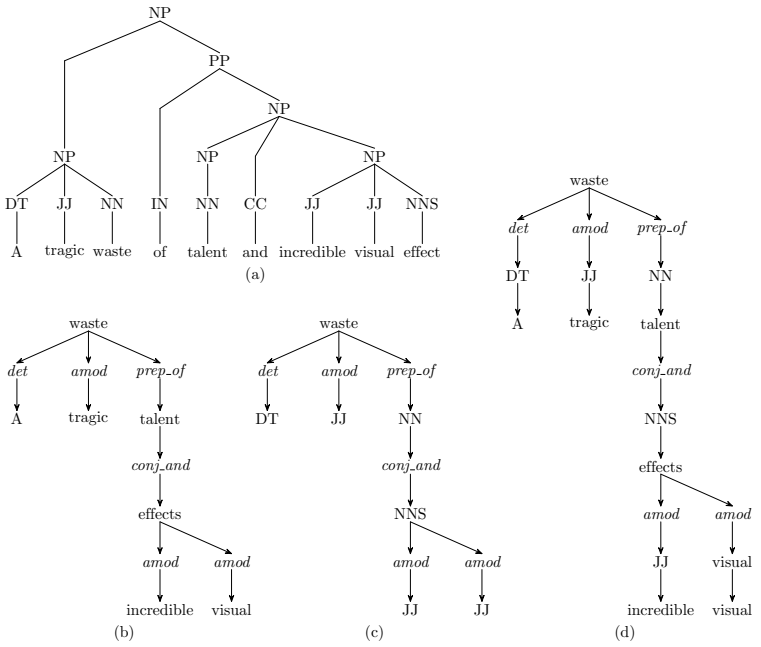

Figure 1: Illustration of the different tree structures employed for convolution kernels. (a) Constituent parse tree (CON); (b) Dependency tree-based words integrated with grammatical relations (DW); (c) Dependency tree in (b) with words substituted by POS tags (DP); (d) Dependency tree in (b) with POS tags inserted before words (DWP).

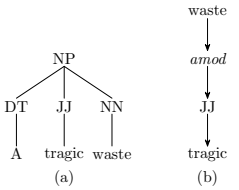

Figure 2: Illustration of the different settings on constituency (CON) and dependency (DWP) parse trees with *tragic* as the indicator word.

Scopes	Trees	Size
Document	32	24
Subjective Sentences	22	27
Constituent Substructures	30	10
Dependency Substructures	40	3

Table 1: The detail of the corpus. Here *Trees* denotes the average number of trees, and *Size* denotes the averaged number of words in each tree.

Unfortunately, in our task there exist several cues indicating the polarity of the document, which are distributed in different sentences. To solve this problem, we define the indicators in this task as subjective words in a polarity lexicon (Wilson et al., 2005). For each polarity indicator, we define the "scope" (the minimal syntactic structure containing at least one subjective word) of each indicator for different representations as follows:

For a constituent tree, a node and its children correspond to a grammatical production. Therefore, considering the terminal node *tragic* in the constituent structure tree in Figure 1(a), we extract the subtree rooted at the grandparent of the terminal, see Figure 2(a). We also use the corresponding sequence of words in the subtree for the sequential kernel.

For a dependency tree, we only consider the subtree containing the lexical items that are directly connected to the subjective word. For instance, given the node *tragic* in Figure 1(d), we will extract its direct parent *waste* integrated with dependency relations and (possibly) POS, as in Figure 2(b).

We further add two *background scopes*, one being subjective sentences (the sentences that contain subjective words), and the entire document.

4 Experiments

4.1 Setup

We carried out experiments on the movie review dataset (Pang and Lee, 2004), which consists of

1000 positive reviews and 1000 negative reviews. To obtain constituency trees, we parsed the document using the Stanford Parser (Klein and Manning, 2003). To obtain dependency trees, we passed the Stanford constituency trees through the Stanford constituency-to-dependency converter (de Marneffe and Manning, 2008).

We exploited Subset Tree (SST) (Collins and Duffy, 2001) and Partial Tree (PT) kernels (Moschitti, 2006) for constituent and dependency parse trees[1], respectively. A sequential kernel is applied for lexical sequences. Kernels were combined using plain (unweighted) summation. Corpus statistics are provided in Table 1.

We use a manually constructed polarity lexicon (Wilson et al., 2005), in which each entry is annotated with its degree of subjectivity (strong, weak), as well as its sentiment polarity (positive, negative and neutral). We only take into account the subjective terms with the degree of strong subjectivity.

We consider two baselines:

- **VK**: bag-of-words features using a *vector kernel* (Pang and Lee, 2004; Ng et al., 2006)
- **Rand**: a number of *randomly selected substructures* similar to the number of extracted substructures defined in Section 3.2

All experiments were carried out using the SVM-Light-TK toolkit[2] with default parameter settings. All results reported are based on 10-fold cross validation.

4.2 Results and Discussions

Table 2 lists the results of the different kernel type combinations. The best performance is obtained by combining VK and DW kernels, gaining a significant improvement of 1.45 point in accuracy. As far as PT kernels are concerned, we find dependency trees with simple words (DW) outperform both dependency trees with POS (DP) and those with both words and POS (DWP). We conjecture that in this case, as syntactic information is already captured by the dependency representation, POS tags can introduce little new information, and will add unnecessary complexity. For example, given the substructure *(waste (amod (JJ (tragic))))*, the PT kernel will use both *(waste (amod (JJ)))* and *(waste (amod (JJ (tragic))))*. We can see that the former is adding no value to the model, as the JJ tag could indicate either positive words (e.g. *good*) or negative words (e.g. *tragic*). In contrast, words are good indicators for sentiment polarity.

Kernels	Doc	Sent	Rand	Sub
VK	87.05			
VK + SW	87.25	86.95	87.25	87.40
VK + SP	87.35	86.95	87.45	87.35
VK + SWP	87.30	87.45	87.30	88.15*
VK + CON	87.45	87.65	87.45	88.30**
VK + DW	87.35	87.50	87.30	**88.50****
VK + DP	87.75*	87.20	87.35	87.75
VK + DWP	87.70*	87.30	87.65	87.80*

Table 2: Results of kernels. Here *Doc* denotes the whole document of the text, *Sent* denotes the sentences that contains subjective terms in the lexicon, *Rand* denotes randomly selected substructures, and *Sub* denotes the substructures defined in Section 3.2. We use "*" and "**" to denote a result is better than baseline VK significantly at $p < 0.05$ and $p < 0.01$ (sign test), respectively.

The results in Table 2 confirm two of our hypotheses. Firstly, it clearly demonstrates the value of incorporating syntactic information into the document-level sentiment classifier, as the tree kernels (CON and D*) generally outperforms vector and sequence kernels (VK and S*). More importantly, it also shows the necessity of extracting appropriate substructures when using convolution kernels in our task: when using the dependency kernel (VK+DW), the result on lexicon guided substructures (Sub) outperforms the results on document, sentence, or randomly selected substructures, with statistical significance ($p<0.05$).

5 Conclusion and Future Work

We studied the impact of syntactic information on document-level sentiment classification using convolution kernels, and reduced the complexity of the kernels by extracting minimal high-impact substructures, guided by a polarity lexicon. Experiments

[1] A SubSet Tree is a structure that satisfies the constraint that grammatical rules cannot be broken, while a Partial Tree is a more general form of substructures obtained by the application of partial production rules of the grammar.

[2] available at http://disi.unitn.it/moschitti/

show that our method outperformed a bag-of-words baseline with a statistically significant gain of 1.45 absolute point in accuracy.

Our research focuses on identifying and using high-impact substructures for convolution kernels in document-level sentiment classification. We expect our method to be complementary with sophisticated methods used in state-of-the-art sentiment classification systems, which is to be explored in future work.

Acknowledgement

The authors were supported by 863 State Key Project No. 2006AA010108, the EuroMatrixPlus FP7 EU project (grant No 231720) and Science Foundation Ireland (Grant No. 07/CE/I1142). Part of the research was done while Zhaopeng Tu was visiting, and Yifan He was at the Centre for Next Generation Localisation (www.cngl.ie), School of Computing, Dublin City University. We thank the anonymous reviewers for their insightful comments. We are also grateful to Junhui Li for his helpful feedback.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

Razvan Bunescu and Raymond Mooney. 2005a. A Shortest Path Dependency Kernel for Relation Extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada, oct. Association for Computational Linguistics.

Razvan Bunescu and Raymond Mooney. 2005b. Subsequence Kernels for Relation Extraction. In Y Weiss, B Sch o lkopf, and J Platt, editors, *Proceedings of the 19th Conference on Neural Information Processing Systems*, pages 171–178, Cambridge, MA. MIT Press.

Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. In *Proceedings of Neural Information Processing Systems*, pages 625–632.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Proceedings of the COLING Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, Manchester, August.

Paul Ferguson, Neil O'Hare, Michael Davy, Adam Bermingham, Paraic Sheridan, Cathal Gurrin, and Alan F. Smeaton. 2009. Exploring the use of paragraph-level annotations for sentiment analysis of financial blogs. In *Proceedings of the Workshop on Opinion Mining and Sentiment Analysis*.

Richard Johansson and Alessandro Moschitti. 2010. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 67–76, Uppsala, Sweden, July.

Mahesh Joshi and Carolyn Penstein-Rose. 2009. Generalizing Dependency Features for Opinion Mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316, Suntec, Singapore, jul. Suntec, Singapore.

Dan Klein and Christopher D Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, jul. Association for Computational Linguistics.

Moshe Koppel and Jonathan Schler. 2005. Using neutral examples for learning polarity. In *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI) 2005*, pages 1616–1616.

Steve Lawrence Kushal Dave and David Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528, ACM. ACM.

Jingjing Liu and Stephanie Seneff. 2009. Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 161–169, Singapore, aug. Singapore.

Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 3518/2005:21–32.

Alessandro Moschitti and Silvia Quarteroni. 2008. Kernels on Linguistic Structures for Answer Extraction. In *Proceedings of ACL-08: HLT, Short Papers*, pages 113–116, Columbus, Ohio, jun. Association for Computational Linguistics.

Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224.

Alessandro Moschitti. 2006. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In *Proceedings of the 17th European Conference on Machine Learning*, pages 318–329, Berlin, Germany,

sep. Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Proceedings.

Vincent Ng, Sajib Dasgupta, and S M Niaz Arifin. 2006. Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 611–618, Sydney, Australia, jul. Sydney, Australia.

Truc-Vien T Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1378–1387.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Barcelona, Spain, jun. Barcelona, Spain.

Michael Wiegand and Dietrich Klakow. 2010. Convolution Kernels for Opinion Holder Extraction. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 795–803, Los Angeles, California, jun. Los Angeles, California.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, oct. Association for Computational Linguistics.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Association for Computational Linguistics. Association for Computational Linguistics.

Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou. 2006. A Composite Kernel to Extract Relations between Entities with Both Flat and Structured Features. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 825–832, Sydney, Australia, jul. Association for Computational Linguistics.