The Use of Topic Representative Words in Text Categorization

Su Nam Kim[♠], Timothy Baldwin[♠♡] and Min-Yen Kan[♣]

[♠]University of Melbourne
[♡]NICTA VRL
[♣]National University of Singapore

December 4, 2009

Introduction

- **Goal**: evaluate the empirical utility of various "topic representative words" for topic classification
- **Motivation**: terms such as keyphrases and named entities are highly indicative of particular topics
- **Question**: can we improve on a simple all-in bag-of-words term representation in the context of text categorization?
- **Significance**: immediate applicability in NLP applications (e.g. text filtering [Amati et al. 1997], WSD [Escudero et al. 2000], automated authorship attribution and genre classification [Diederich et al. 2003])

Related Work

- Different learners: naive Bayes (NB), Rocchio, Decision trees (DT), SVMs (Dumais et al. 1998, Yang and Liu 1997, Joachims 1998)
- Different term representations: n-grams (Cavnar and Trenkle 1994), clustered words (Barker and MacCallum 1998), complex nominals (Moschitti and Basili 2004), important sentences (Mihalcea and Hassan 2005), keyphrases (Hulth and Megayesi 2006), ...
- Different term weights: mutual information (Lewis 1992), chi-square (Yang and Pedersen 1997), gain ratio (Debole and Sebastiani 2003), ...

Topic Representative Words

- Zone-based terms: previous research (Mihalcea and Hassan 2005, Nguyen and Kan 2007) has shown that sentences in particular "zones" (e.g. title or first sentence) contain more keyphrases

- Keyphrases: keyphrases are sets of words that capture the topic of the document

- Domain-Specific Words (DSW): domain-variant of $TF{\cdot}IDF$ (e.g. $goods \rightarrow$ "trade")

- Named Entities (NEs): NEs often associated with particular domains (e.g. $Gulf, Kuwait \rightarrow$ "oil")

Zone-based Terms

- Term extraction methodology:
  - ⋆ 1-grams from titles
  - ⋆ 1-grams from first sentences, as they tend to contain more information (Mihalcea and Hassan 2005)
  - ⋆ Data: subset of Reuter-21578 containing 90 domains

Type	F1(≥1)	F2(≥2)	F3(≥3)
Title words	8,622	3,878	2,357
First sentence words	11,565	5,819	3,905

Keyphrases

- Background
  - ⋆ condensed summary of the document and high-quality index terms
  - ⋆ a large body of study done using (a) document cohesion (b) keyphrase cohesion, and (c) term cohesion
- Extraction: scoring using $TF{\cdot}IDF$ and relative position of words (KEA features), then select top-N candidates according to 3 thresholds

$$Score = TF{\cdot}IDF + (1 - \frac{first\ position\ of\ W_i}{\#\ of\ total\ terms})$$

- Statistics: count 1-grams as well as NPs (w/ NPs, count individual 1-grams) $\rightarrow$ 1+NP

Length	T1(.02)	T2(.04)	T3(.06)
original	7,889	5,733	4,497
1+NP	25,343	15,257	10,679

- Performance on Keyphrase Extraction: with 100 sample documents

	Precision	Recall	Fscore
T1	9.76%	23.85%	13.85%
T2	15.32%	15.62%	15.47%
T3	21.02%	10.86%	14.32%

Domain-Specific Words

- Background:
  - ⋆ word-sense based vs. document statistics
  - ⋆ traditionally supervised, based on large corpus (Rigutini et al. 2006), cohesion or frequency (Drouin 2004, Park et al. 2008)

- Extraction:
  - ⋆ our proposed method (**D1**)

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

$$IDF_i = log(\frac{|D|}{|\{d : t_i \in d\}|})$$

- ⋆ Park et al. (2008) (**D2**)

$$domain\_specificity(w) = \frac{\frac{c_d(w)}{N_d}}{\frac{c_g(w)}{N_g}}$$

- DSW verification
  - ⋆ over 23 domains which have at least 5 articles in both test and training data sets
  - ⋆ accuracy: 40.6% vs. 36.6% for D1 and D2, respectively

- Statistics:

Method	Length	T1	T2	T3
D1	original	2,918	1,573	1,157
	1+NP	3,969	1,918	1,344
D2	original	3,692	2,759	2,368
	1+NP	7,169	5,021	4,215

	Overlap	D1	D2
T1	1,612	55.24%	43.67%
T2	593	37.70%	21.49%
T3	404	34.92%	17.06%

Named Entities

- Background:
  - ⋆ basic approach: unsupervised, supervised or semi-supervised approach using models such as hidden Markov models (HMMs) or conditional random fields (CRFs)
  - ⋆ shown to enhance Question-Answering (Molla et al. 2006), web search (Sekine et al. 2002)
- Extraction:
  - ⋆ using NER toolkit developed by UIUC
  - ⋆ four entities (i.e. PER, LOC, ORG, MISC)

- Statistics:

Length	F1($\geq$1)	F2($\geq$2)	F3($\geq$3)
original	11,431	6,538	4,650
1+NP	23,440	9,883	6,234

Data and Experimental Setup

- Data Collection
  - ⋆ Modified Lewis split from Reuter collection
  - ⋆ 7,771 training and 3,019 test documents over 90 categories/domains (topic categorization task)
- Experimental Setup
  - ⋆ Preprocessing: POS tagging, lemmatization, $TF{\cdot}IDF$ term weighting
  - ⋆ Learner: SVM
  - ⋆ Baseline: using 1-gram with F3 (**B3**) $\rightarrow$ micro-average F-score, 78.54%

Topic Categorization Results

Word	Length	T1/F1	T2/F2	T3/F3
Baseline	1	77.80%	78.09%	78.54%
Title(T)	1	78.09%	78.18%	78.18%
First(F)	1	78.18%	78.09%	77.98%
Keyphrase(K)	1	78.57%	78.07%	78.27%
	1+NP	78.36%	78.24%	78.24%
Domain(D1)	1	77.00%	76.50%	74.49%
	1+NP	77.00%	76.50%	74.49%
Domain(D2)	1	75.58%	73.90%	72.98%
	1+NP	75.58%	73.90%	72.98%
NE(N)	1	76.91%	76.35%	75.76%
	1+NP	77.06%	76.35%	76.03%
T+F+K+D1+N	1	78.54%	78.48%	78.36%
	1+NP	78.66%	78.30%	78.48%
T+F+K+D2+N	1	78.60%	78.51%	78.57 %
	1+NP	78.69%	78.63%	78.77%

Word	Length	T1/F1	T2/F2	T3/F3
Baseline	1	77.80%	78.09%	78.54%
B3+Title	1	78.30%	78.42%	78.15%
B3+First	1	78.36%	78.21%	78.39%
B3+Keyphrase	1	78.72%	78.42%	78.60%
	1+NP	78.83%	**78.89%**	78.69%
B3+Domain(D1)	1	78.51%	78.63%	78.51%
	1+NP	78.51%	78.63%	78.51%
B3+Domain(D2)	1	78.07%	77.95%	78.27%
	1+NP	78.07%	77.95%	78.27%
B3+NE	1	78.18%	78.27%	78.54%
	1+NP	78.18%	78.24%	78.07%
B3+T+F+K+D1+N	1	78.80%	78.83%	78.77%
	1+NP	**78.95%**	78.69%	78.75%
B3+T+F+K+D2+N	1	78.83%	78.80%	**78.98%**
	1+NP	**78.95%**	**78.89%**	**78.98%**

Performance on Top-10 Topics

Feature sets	F-score
Baseline	89.55%
Individual	89.59%
Individual+1-gram	89.96%
All candidates	90.02%
All candidates+1-gram	90.07%

Future Work and Summary

- **Future Work**
  - ⋆ achieve higher performance on keyphrase extraction
  - ⋆ investigate more reliable method to extract domain-specific words

- **Individual candidates**
  - ⋆ only keyphrases outperformed baseline
  - ⋆ w.r.t. frequencies, words w/ locality, keyphrases performed better than baseline
  - ⋆ considering the small amount of words, domain-specific words and NEs performed well

- **Combined features**
  - ⋆ only keyphrases w/ BoW outperformed baselines (similar to performance of individual methods)

- **1-gram vs. 1+NP** results indicate that the added NPs produced a slight improvement in results (cf. Hulth and Megayesi (2006))

- **Our method vs. Park et al. 2008** using DSW collected by our method performed better

References

[1] G. Amati and D. DAloisi and V. Giannini and F. Ubaldini, A framework for filtering news and managing distributed data, Journal of Universal Computer Science, 1997, 3(8), pp. 1007–1021.

[2] L.D. Barker and A.K. McCallum, Distributional clustering of words for text categorization, In Proceedings of 21st ACM International Conference on Research and Development in Informatoin Retrieval, 1998, pp.96–103.

[3] K. Barker and N. Corrnacchia, Using noun phrase heads to extract document keyphrases, In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, 2000, pp. 40–52.

[4] W.B. Cavnar and J.M. Trenkle, N-gram-based text categorization, In Proceedings of SDAIR, 1994, pp. 161–175.

[5] M. Collins and Y. Singer, Unsupervised Models for Named Entity Classification, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999, pp. 100–110.

[6] D. Okanohara and Y. Miyao and Y. Tsuruoka and J. Tsujii, Improving the Scalability of Semi-Markov Conditional Random Fields for Named Entity Recognition, In Proceedings of COLING/ACL, 2006, pp. 465–472.

[7] F. Debole and F. Sebastiani, Supervised term weighting for automated text categorization, In 18th ACM Symposium on Applied Computing, 2003, pp.784–788.

[8] J. Diederich and J. Kindermann and E. Leopold and G. Paass, Authorship attribution with support vector machines, Applied Intelligence, 2003, 19(1/2), pp.109–123.

[9] P. Drouin, Detection of Domain Specific Terminology Using Corpora Comparison, In Proceedings of the 4th LREC, 2004, pp. 79–82.

[10] S. Dumais and J. Platt and D. Heckerman and M. Sahami, Inductive learning algorithms and representations for text categorization, In Proceedings of CIKM, 1998, pp. 148–155.

[11] G. Escudero and L. Marquez and G. Rigau, Boosting applied to word sense disambiguation, In Proceedings of 11th European Conference on Machine Learning, 2000, pp. 129–141.

[12] E. Frank and G.W. Paynter and I. Witten and C. Gutwin and C.G. Nevill-Manning, Domain Specific Keyphrase Extraction, In Proceedings of the 16th IJCAI, 1999, pp. 668–673.

[13] A. Hulth and B. Megayesi, A Study on Automatically Extracted Keywords in Text Categorization, In Proceedings of the 21st COLING/ACL, 2006, pp. 537–544.

[14] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, In Proceedings of ECML, 1998, pp. 137–142.

[15] M. Kida, M. Tonoike, T. Utsuro and S. Sato, Domain Classification of Technical Terms Using the Web, Systems and Computers, 2007, 38(14), pp. 2470–2482.

[16] S. Kim, T. Baldwin and M-Y. Kan, An Unsupervised Approach to Domain-Specific Term Extraction, In Proceedings of the Australasian Language Technology Workshop 2009, to appear.

[17] Y. Ko and J. Park and J. Seo, Improving text categorization using the importance of sentences, Information Processing and Management, 2004, 40(1), pp. 65–79.

[18] D.D. Lewis, An evaluation of phrasal and clustered representations on a text categorization task, In 15th ACM International Conference on Research and Development in Informaton Retrieval, 1992, pp. 37–50.

[19] B. Magnini and C. Strapparava and G. Pezzulo and A. Gliozzo, The role of domain information in word sense disambiguation, Natural Language Engineering, 2002, 8(4), pp. 359–373.

[20] Y. Matsuo and M. Ishizuka, Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, International Journal on Artificial Intelligence Tools, 2004, 13(1), pp. 157–169.

[21] R. Mihalcea and S. Hassan, Using the essence of texts to improve document classification, In Proceedings of RANLP, 2005.

[22] G. Minnen and J. Carroll and D. Pearce, Applied morphological processing of English, Natural Language Engineering, 2001, 7(3), pp. 207–223.

[23] D. Molla and M. van Zaanen and D. Smith, Named Entity Recognition for Question Answering, In Proceedings of ALTW, 2006, pp. 51–58.

[24] A. Moschitti and R. Basili, Complex linguistic features for text classification, In Proceedings of 26th European Conference on Information Retrieval Research, 2004, pp.181–196.

[25] D. Nadeau and P.D. Turney and S. Matwin, Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity, In cogprints, 2006, pp. 266–277.

[26] T. Nguyen and M.Y. Kan, Key phrase Extraction in Scientific Publications, In Proceeding of International Conference on Asian Digital Libraries, 2007, pp. 317-326.

[27] S. Pakhomov, Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts, In Proceedings of 40th ACL, 2002, pp. 160–167.

[28] Y. Park and R.J. Byrd and B. Boguraev, Automatic Glossary Extraction Beyond Terminology Identification, In Proceedings of COLING, 2004, pp. 48–55.

[29] Y. Park and S. Patwardhan and K. Visweswariah and S.C. Gates, An Empirical Analysis of Word Error Rate and Keyword Error Rate, In Proceedings of International Conference on Spoken Language Processing, 2008, pp. 2070–2073.

[30] L. Ratinov and D. Roth, External Knowledge and Non-local Features in Named Entity Recognition, In Proceedings of NAACL, 2009.

[31] L. Rigutini and E. Di Iorio and M. Ernandes and M. Maggini, Automatic term categorization by extracting knowledge from the Web, In Proceedings of 17th ECAI, 2006, pp. 531–535.

[32] G. Salton and A. Wong and C.S. Yang, A vector space model for automatic indexing, Communications of the ACM, 1975, 18(11), pp. 61–620.

[33] F. Sebastiani, Machine learning in automated text categorization, ACM Computering Surveys, 2002, 34(1), pp. 1–47.

[34] S. Sekine and K. Sudo and C. Nobata, Extended Named Entity Hierarchy, In Proceedings of LREC, 2002.

[35] T. Tomokiyo and M. Hurst, A Langauge Model Approach to Keyphrase Extraction, In Proceedings of ACL Workshop on Multiword Expressions, 2003, pp.33–40.

[36] P. Turney, Learning to Extract Keyphrases from Text, In National Research Council, Institute for Information Technology, Technical Report ERB-1057, 1999.

[37] P. Turney, Coherent keyphrase extraction via Web mining, In Proceedings of the 18th IJCAI, 2003, pp. 434–439.

[38] X. Wan and J. Xiao, CollabRank: towards a collaborative approach to single-document keyphrase extraction, In Proceedings of COLING, 2008, pp. 969–976.

[39] I. Witten and G. Paynter and E. Frank and C. Gutwin and G. Nevill-Manning, KEA:Practical Automatic Key phrase Extraction, In Proceedings of the fourth ACM conference on Digital libraries, 1999, pp.254–256.

[40] Y. Yang and X. Liu, A re-examination of text categorization methods, In Proceedings of SIGIR, 1997, pp. 42–49.

[41] Y. Yang and J.O. Pedersen, A comparative study on feature selection in text categorization, In Proceedings of 14th International Conference on Machine Learning, 1997, pp. 412–420.