Supervised Keyphrase Extraction as Positive Unlabeled Learning

Lucas Sterckx, Thomas Demeester, Chris Develder
Ghent University - iMinds
Ghent, Belgium
{lusterck,tdmeeste,cdvelder}@intec.ugent.be

Cornelia Caragea
University of North Texas
Texas, USA
Cornelia.Caragea@unt.edu

Abstract

The problem of noisy and unbalanced training data for supervised keyphrase extraction results from the subjectivity of keyphrase assignment, which we quantify by crowdsourcing keyphrases for news and fashion magazine articles with many annotators per document. We show that annotators exhibit substantial disagreement, meaning that single annotator data could lead to very different training sets for supervised keyphrase extractors. Thus, annotations from single authors or readers lead to noisy training data and poor extraction performance of the resulting supervised extractor. We provide a simple but effective solution to still work with such data by reweighting the importance of unlabeled candidate phrases in a two stage Positive Unlabeled Learning setting. We show that performance of trained keyphrase extractors approximates a classifier trained on articles labeled by multiple annotators, leading to higher average F_1scores and better rankings of keyphrases. We apply this strategy to a variety of test collections from different backgrounds and show improvements over strong baseline models.

1 Introduction

Keyphrase extraction is the task of extracting a selection of phrases from a text document to concisely summarize its contents. Applications of keyphrases range from summarization (D'Avanzo et al., 2004) to contextual advertisement (Yih et al., 2006) or simply as aid for navigation through large text corpora.

Existing work on automatic keyphrase extraction can be divided in supervised and unsupervised approaches. While unsupervised approaches are domain independent and do not require labeled training data, supervised keyphrase extraction allows for more expressive feature design and is reported to outperform unsupervised methods on many occasions (Kim et al., 2012; Caragea et al., 2014). A requirement for supervised keyphrase extractors is the availability of labeled training data. In literature, training collections for supervised keyphrase extraction are generated in different settings. Keyphrases for text documents are either supplied by the authors or its readers. In the first case, authors of academic papers or news articles assign keyphrases to their content to enable fast indexing or to allow for the discovery of their work in electronic libraries (Frank et al., 1999; Hulth, 2003; Bulgarov and Caragea, 2015). Other collections are created by crowdsourcing (Marujo et al., 2012) or decided after deliberation by a small group of readers (Wan and Xiao, 2008). A minority of test collections provide multiple opinions per document, but even then the amount of opinions per document is kept minimal (Nguyen and Kan, 2007).

Supervision for keyphrase extraction faces several shortcomings. Candidate phrases (generated in a separate candidate generation procedure), which are not annotated as keyphrases, are seen as non-keyphrase and are used as negative training data for the supervised classifiers. First, on many occasions these negative phrases outnumber true keyphrases many times, creating an unbalanced training set (Frank et al., 1999; Kim et al., 2012). Second, as Frank et al. (1999) noted: authors do not always choose keyphrases that best describe the

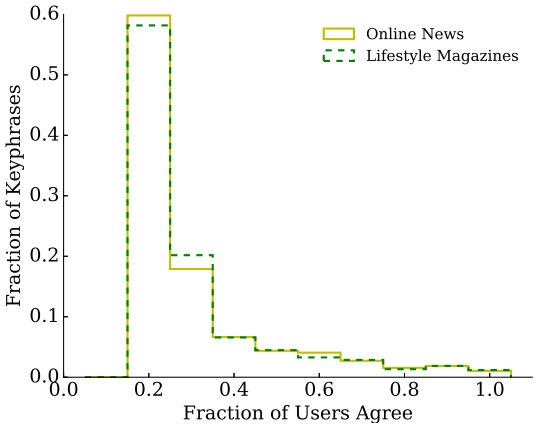

Figure 1: This plot shows the fraction of all keyphrases from the training set agreed upon versus the fraction of all annotators.

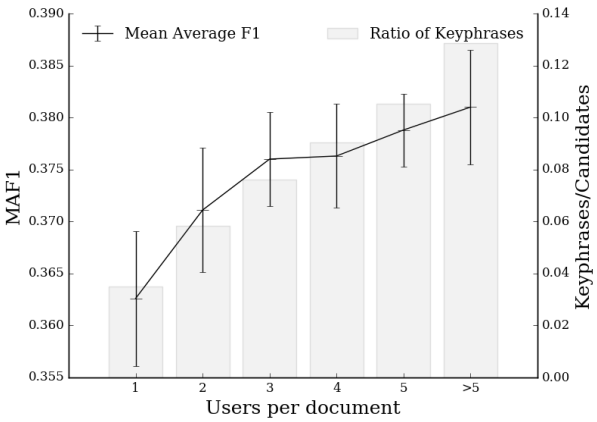

Figure 2: Effect of overlap on extraction performance.

content of their paper, but they may choose phrases to slant their work a certain way, or to maximize its chance of being noticed by searchers. Another problem is that keyphrases are inherently subjective, i.e., keyphrases assigned by one annotator are not the only correct ones (Nguyen and Kan, 2007). These assumptions have consequences for training, developing and evaluating supervised models. Unfortunately, a large collection of annotated documents by reliable annotators with high overlap per document is missing, making it difficult to study disagreement between annotators, the resulting influence on trained extractors, as well as to provide a reliable evaluation setting. In this paper, we address these problems by creating a large test collection of articles with many different opinions per article, evaluate the effect on extraction performance and present a procedure for supervised keyphrase extraction with noisy labels.

2 Noisy Training Data for Supervised Keyphrase Extraction

A collection of online news articles and lifestyle magazine articles was presented to a panel of 357 annotators of various ages and backgrounds, who were trained to select a *limited number* of *short* phrases that concisely summarize the documents' contents. Each document was presented multiple times to different users. Each user was assigned with 140 articles, but was not required to finish the full assignment. The constructed training collections have on average six and up to ten different opinions per each article. We visualize the agreement on single keyphrases in Fig. 1, which shows the fraction of annotated keyphrases versus agreement by the complete set of readers. Agreement on keyphrases is low, a large fraction of all assigned keyphrases (>50%) are only assigned by single annotators. We note that different sets of keyphrases by different annotators are the result of the subjectiveness of the task, of different interpretations by the annotators of the document, but also because of semantically equivalent keyphrases being annotated in different forms, e.g., "Louis Michel" vs. "Prime Minister Louis Michel" or "Traffic Collision" vs. "Car Accident". The observation in Figure 1 has important consequences for training models on keyphrases annotated by a single annotator, since other annotators may have chosen some among the ones that the single selected annotator did not indicate (and hence these should not be used as negative training data). A single annotator assigning keyphrases to 100 documents results on average in a training set with 369 positive training instances and 4,981 negative training instances generated by the candidate extractor. When assigning these 100 documents to 9 other annotators, the amount of positive instances increases to 1,258 keyphrases, which means that labels for 889 keyphrase candidates, or 17% of the original negative candidates when training on annotations by a single annotator, can be considered noise and re-

labeled. As a result, ratios of positive to negative data also change drastically. We visualize the effect of overlap of annotations in Figure 2. Classifiers trained on the aggregated training collection with multiple opinions (using all assigned keyphrases at least once as positive training data) perform better on held-out test collections with keyphrases of high agreement (assigned by > 2 annotators). When increasing overlap, the amount of positive candidates increases and as a result, the Mean Average F_1 of the corresponding classifier. We detail our experimental setup and supervised classifier in Section 4.

3 Reweighting Keyphrase Candidates

Observations described in Section 2 indicate that unlabeled keyphrase candidates are not reliable as negative examples by default. A more suitable assumption is to treat supervised keyphrase extraction as Positive Unlabeled Learning, i.e., an incomplete set of positive examples is available and a set of unlabeled examples of which some are positive and others negative. This topic has received much attention as it knows many applications (Ren et al., 2014; du Plessis et al., 2014), but has not been linked to supervised keyphrase extraction. We base our approach on work by Elkan and Noto (2008) and modify the supervised extractor by assigning individual weights to training examples, and instead of assuming the noise is random, we assign weights depending on the document and the candidate.

By reweighting importance of training samples, we seek to model the annotations by multiple annotators and the uncertainty of negative keyphrase candidates, using annotations by a single annotator. In a first stage, we train a classifier on the single annotator data and use predictions on the negative or unlabeled candidates, to reweigh training instances. The reweighted training collection is then used to train a second classifier to predict a final ranking or labels of the keyphrase candidates.

Positive examples are given unit weight and unlabeled examples are duplicated; one copy of each unlabeled keyphrase candidate x is made positive with weight $w(x) = P(keyphrase|x, s = 0)$ and the other copy is made negative with weight $1 - w(x)$ with s indicating whether x is labeled or not. The new estimate $E[h]$ of a keyphrase classifier function $h(x, keyphrase)$ then becomes,

$$\frac{1}{m}\left(\sum_{(x,s=1)} h(x, keyphrase) + \sum_{(x,s=0)} w(x)h(x, keyphrase) + (1 - w(x))\, h(x, \neg keyphrase)\right) \quad (1)$$

where m is the training set size.

Instead of assigning this weight as a constant factor of the predictions by the initial classifier as in Elkan and Noto (2008), two modifications showed to improve the weight estimate, $w(x) \leq 1$. We normalize probabilities $P(keyphrase, x, s = 0)$ to candidates not included in the initial set of keyphrases per document. Next to this self-predicted probability, we include a simple measure indicating pairwise coreference between unlabeled candidates and known keyphrases in a function Coref$(candidate, keyphrase) \in \{0, 1\}$, returning 1 if one of the binary indicator features, presented in (Bengtson and Roth, 2008) and shown in Table 1, is present. The self-predicted probability is summed with the output of the coreference resolver and the final weight becomes:

$$w(x) = min\left(1, \frac{P(keyphrase|x)}{\max_{(x',s=0)\in d} P(keyphrase|x')} \cdot + \max_{keyphrase\in d} \text{Coref}(x, keyphrase)\right) \quad (2)$$

with d being a document from the training collection.

Feature	Definition
Head match	$head_{keyphrase} == head_{candidate}$
Extent match	$extent_{keyphrase} == extent_{candidate}$
Substring	$head_{keyphrase}$ substring of $head_{candidate}$
Alias	$acronym(head_{keyphrase}) == head_{candidate}$

Table 1: String relation features for coreference resolution

4 Experiments and Results

Hasan and Ng (2010) have shown that techniques for keyphrase extraction are inconsistent and need to be tested across different test collections. Next to our collections with multiple opinions (*Online News* and *Lifestyle Magazines*), we apply the reweighting strategy on test collections with sets of author-assigned keyphrases: two sets from CiteSeer abstracts from the World Wide Web Conference

	Test Collections				
Name	Online News	Lifestyle Magazines	WWW	KDD	Inspec
Type	Sports Articles	Fashion, Lifestyle	WWW Paper Abstracts	KDD Paper Abstracts	Paper Abstracts
# Documents	1,259	2,202	1,895	1,011	500
# Keyphrases	19,340	29,970	3,922	1,966	4,913
⊘ Keyphrases/User	5.7	4.7	/	/	/
⊘ Keyphrases/Document	15.4	13.7	2.0	1.8	9.8
⊘ Tokens/Document	332	284	164	195	134
⊘ Candidate Keyphrases/Doc.	52	49	47	54	34
1/2/3/3+ -gram distribution (%)	55/27/9/9	58/25/9/8	63/27/8/2	60/28/9/3	13/53/25/9

Table 2: Description of test collections.

Method	Online News		Lifestyle Magazines		WWW		KDD		Inspec	
	MAF_1	P@5	MAF_1	P@5	MAF_1	P@5	MAF_1	P@5	MAF_1	P@5
Single Annotator	.364	.416	.294	.315	.230	.189	.266	.200	.397	.432
Multiple Annotators	.381	.426	.303	.327	/	/	/	/	/	/
Self Training	.366	.417	.301	.317	.236	.190	.269	.196	.401	**.434**
Reweighting (Elkan and Noto, 2008)	.364	.417	.297	.313	.238	.189	**.275**	**.201**	.401	.429
Reweighting +Norm +Coref	**.374**	**.419**	**.305**	**.322**	**.245**	**.194**	**.275**	.200	**.402**	**.434**

Table 3: Mean average F_1score per document and precision for five most confident keyphrases on different test collections.

(*WWW*) and Knowledge Discovery and Data Mining (*KDD*), similar to the ones used in (Bulgarov and Caragea, 2015). The *Inspec* dataset is a collection of 2,000 abstracts commonly used in keyphrase extraction literature, where we use the ground truth phrases from controlled vocabulary (Hulth, 2003). Descriptive statistics of these test collections are given in Table 2.

We use a rich feature set consisting of statistical, structural, and semantic properties for each candidate phrase, that have been reported as effective in previous studies on supervised extractors (Frank et al., 1999; Hulth, 2003; Kim and Kan, 2009): (i) term frequency, (ii) number of tokens in the phrase, (iii) length of the longest term in the phrase, (iv) number of capital letters in the phrase, (v) the phrase's POS-tags, (vi) relative position of first occurrence, (vii) span (relative last occurrence minus relative first occurrence), (viii) TF*IDF (IDF's trained on large background collections from the same source) and (ix) Topical Word Importance, a feature measuring the similarity between the word-topic topic-document distributions presented in (Sterckx et al., 2015), with topic models trained on background collections from a corresponding source of content.

As classifier we use gradient boosted decision trees implemented in the XGBoost package (Chen and Guestrin, 2016). During developing, this classifier consistently outperformed Naive Bayes and linear classifiers like logistic regression or support vector machines.

We compare the reweighting strategy with uniform reweighting and strategies to counter the imbalance or noise of the training collections, such as subsampling, weighting unlabeled training data as in (Elkan and Noto, 2008), and self-training in which only confident initial predictions are used as positive and negative data. For every method, global thresholds are chosen to optimize the macro averaged F_1per document (MAF_1). Next to the threshold sensitive F_1, we report on ranking quality using the Precision@5 metric.

Results are shown in Table 3 with five-fold cross-validation. To study the effect of reweighting, we limit training collections during folds to 100 documents for each test collection. Our approach consistently improves on single annotator trained classifiers, on one occasion even outperforming a training collection with multiple opinions. Compensating for imbalance and noise tends to have less effect when the ratio of keyphrases versus candidates is high (as for *Inspec*) or training collection is very large. When the amount of training documents increases, the ratio of noisy versus true negative labels drops.

5 Conclusion

It has been suggested that keyphrase annotation is highly subjective. We present two data sets where

we purposely gathered multiple annotations of the same document, as to quantify the limited overlap between keyphrases selected by different annotators. We suggest to treat non-selected phrases as *unlabeled* rather than *negative* training data. We further show that using multiple annotations leads to more robust automatic keyphrase extractors, and propose reweighting of single annotator labels based on probabilities from a first-stage classifier. This reweighting approach outperforms other single-annotator state-of-the-art baseline automatic keyphrase extractors on different test collections, when we normalize probabilities per document and include co-reference indicators.

References

[Bengtson and Roth2008] Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 294–303.

[Bulgarov and Caragea2015] Florin Adrian Bulgarov and Cornelia Caragea. 2015. A comparison of supervised keyphrase extraction models. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 13–14.

[Caragea et al.2014] Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar, October. Association for Computational Linguistics.

[Chen and Guestrin2016] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

[D'Avanzo et al.2004] Ernesto D'Avanzo, Bernardo Magnini, and Alessandro Vallin. 2004. Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004. In *Proceedings of the 2004 DUC*.

[du Plessis et al.2014] Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. 2014. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 703–711.

[Elkan and Noto2008] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 213–220.

[Frank et al.1999] Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-manning. 1999. Domain specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on AI*, pages 668–673.

[Hasan and Ng2010] Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art. In *Proceedings of the 23rd COLING*, COLING 2010, pages 365–373, Stroudsburg, PA, USA.

[Hulth2003] Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223.

[Kim and Kan2009] Su Nam Kim and Min-Yen Kan. 2009. Re-examining automatic keyphrase extraction approaches in scientific articles. In *Proceedings of the workshop on multiword expressions: Identification, interpretation, disambiguation and applications*, pages 9–16. Association for Computational Linguistics.

[Kim et al.2012] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2012. Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation*, 47(3):723–742, December.

[Marujo et al.2012] Luis Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, and Jo ao P. Neto. 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proceedings of LREC 2012*. ELRA.

[Nguyen and Kan2007] Thuy Dung Nguyen and Min-Yen Kan. 2007. Key phrase extraction in scientific publications. In *Proceeding of International Conference on Asian Digital Libraries*, pages 317–326.

[Ren et al.2014] Yafeng Ren, Donghong Ji, and Hongbin Zhang. 2014. Positive unlabeled learning for deceptive reviews detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 488–498.

[Sterckx et al.2015] Lucas Sterckx, Thomas Demeester, Johannes Deleu, and Chris Develder. 2015. Topical word importance for fast keyphrase extraction. In *Proceedings of the 24th International Conference on*

World Wide Web Companion, pages 121–122. International World Wide Web Conferences Steering Committee.

[Wan and Xiao2008] Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI 2008, pages 855–860.

[Yih et al.2006] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. 2006. Finding advertising keywords on web pages. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 213–222, New York, NY, USA. ACM.