Supervised Keyphrase Extraction as Positive Unlabeled Learning

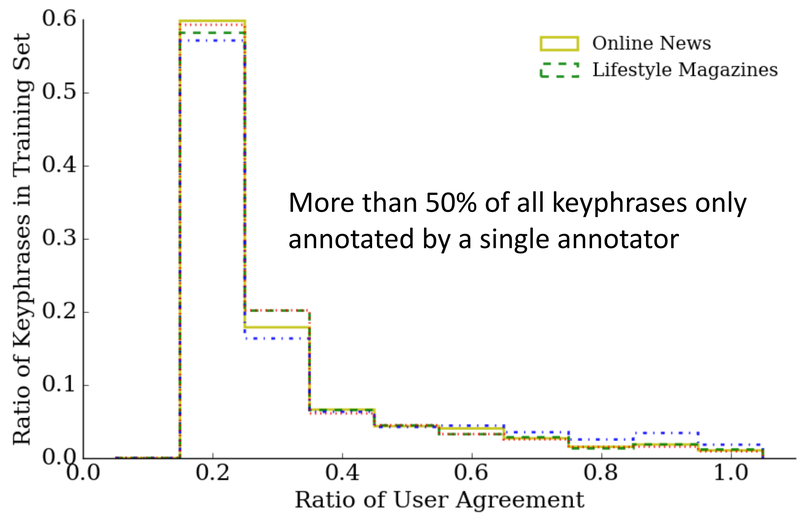
Lucas Sterckx*, **Cornelia Caragea†**, **Thomas Demeester***, Chris Develder*
*Ghent University – imec, †University of North Texas

Training Data for Supervised Keyphrase Extraction

- **Supervised keyphrase extraction** = binary classification of keyphrase candidates
  - State-of-the-art but requires **training data**
- Problems with keyphrase annotations and resulting training data
  - **Noisy :** keyphrases are highly subjective
  - **Unbalanced** : negatives outweigh positives many times

We create large test collections of articles with many different opinions per document, evaluate the effect on extraction performance, and present a procedure for supervised keyphrase extraction with noisy labels.

Disagreement on Keyphrases



Keyphrase Extraction as Positive Unlabeled Learning

Document

After **10 years** and a **journey of four billion miles**, **the European Space Agency**'s **Rosetta spacecraft** arrived at its destination on **Wednesday** for the first extended, close examination of a **comet**.
A **six-minute thruster** firing at 5 a.m. Eastern time, the last in a series of 10 over the past few months, slowed **Rosetta** to the pace of a person walking, about two miles per hour relative to the speed of its target,**Comet 67P/Churyumov-Gerasimenko**.

Single Annotator

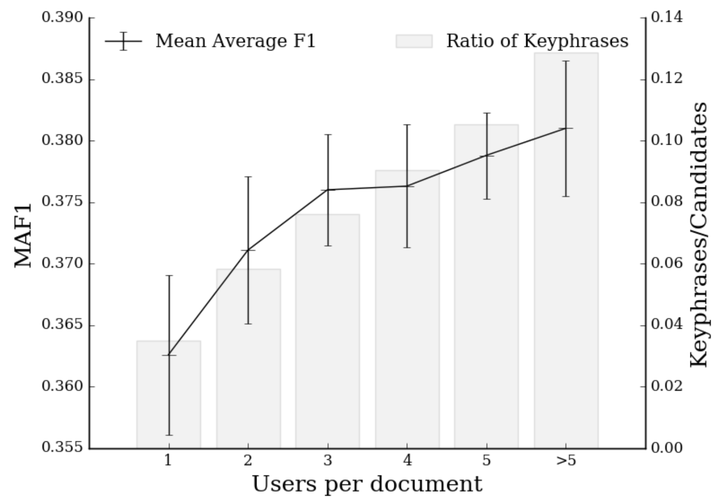
	Keyphrase
+	Rosetta spacecraft
+	Comet 67P/Churyumov-Gerasimenko
-	European Space Agency
-	Journey of four billion miles
-	Six-minute thruster
-	at 5 a.m. Eastern time

Multiple Annotators

	Keyphrase
+	Rosetta spacecraft
+	Comet 67P/Churyumov-Gerasimenko
+	European Space Agency
+	Journey of four billion miles
+	Six-minute thruster
-	at 5 a.m. Eastern time

\+

Overlap by multiple annotators results in better training data for supervised extractor



Positive Unlabeled Learning

Negative Training data becomes weighted Positive **and** Negative (Elkan, 2008)

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

Use probabilities by initial classifier as weights,
normalize per document and add coreference indicator

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

+ / -	Keyphrase
1.00/0.00	Rosetta spacecraft
1.00/0.00	European Space Agency
0.90/0.10	Comet 67P/Churyumov-Gerasimenko
0.65/0.35	Journey of four billion miles
0.45/0.55	Six-minute thruster
0.03/0.97	at 5 a.m. Eastern time

Experiments

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	.364	.417	.297	.313	.238	.189	.275	.201	401	.429
Reweighting + Norm. + Coref	.374	.419	.305	.322	.245	.194	.275	.200	.402	.434

Conclusions

- Keyphrase datasets with multiple annotations per document
- Supervised keyphrase extraction as **Positive Unlabeled Learning**
  - Treat non-selected phrases as **unlabeled instead of negative**
  - Reweigh keyphrases using noisy classifier prediction, **normalization and coreference**
- Future work :
  Evaluation measures for keyphrases

lusterck@intec.ugent.be
@lusterck
lusterck.github.io

emnlp2016