GHENT UNIVERSITY | FACULTY OF ENGINEERING AND ARCHITECTURE

UNT UNIVERSITY OF NORTH TEXAS

Supervised Keyphrase Extraction as Positive Unlabeled Learning

Lucas Sterckx†, **Cornelia Caragea***, **Thomas Demeester**†, Chris Develder†

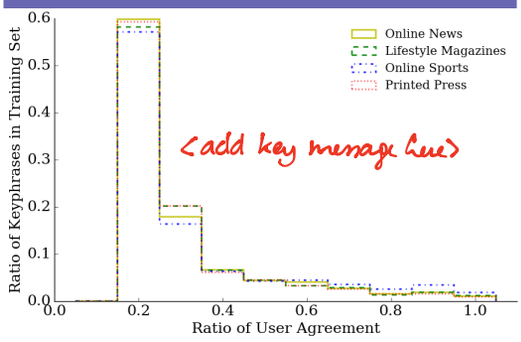
†Ghent University – imec, *University of North Texas

Training Data for Supervised Keyphrase Extraction

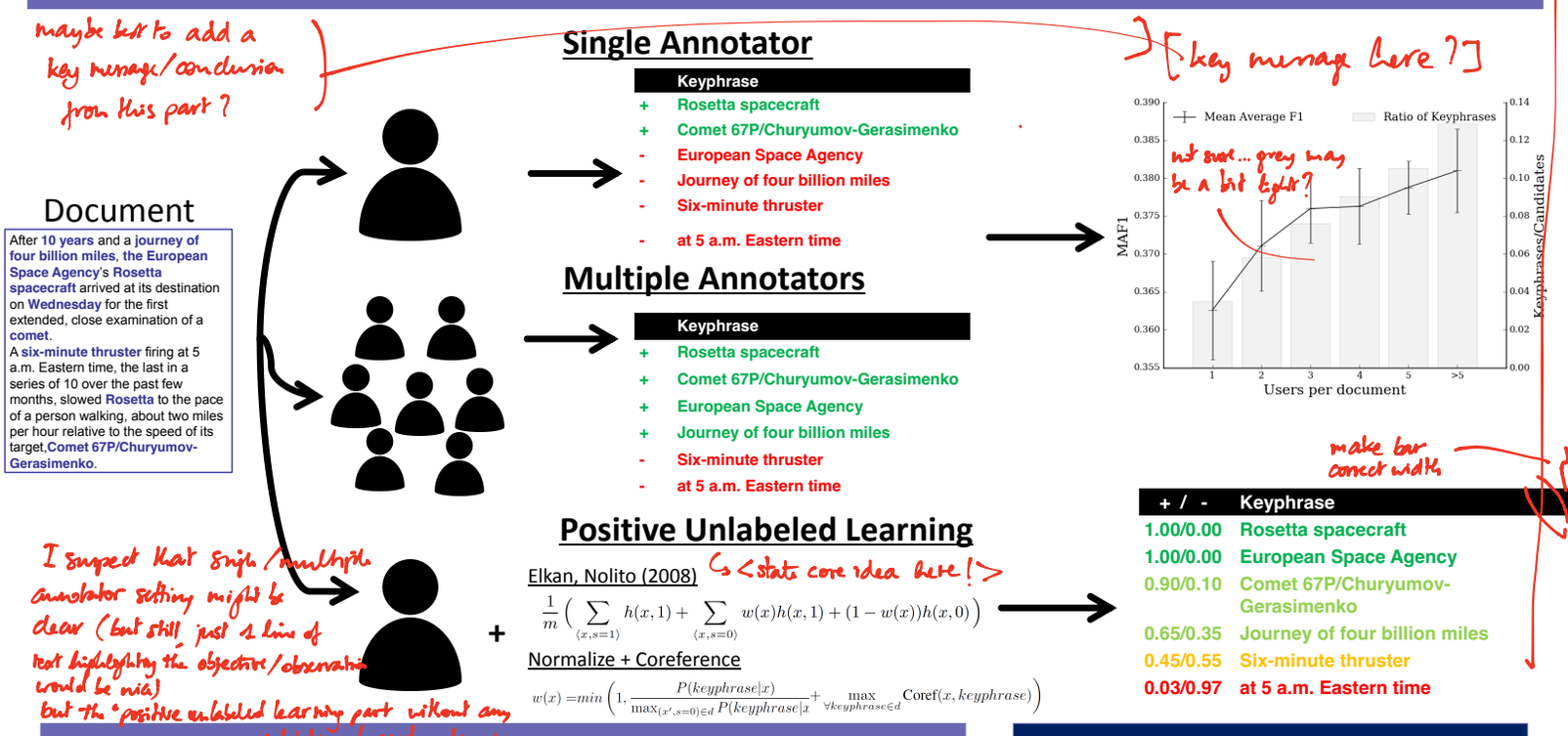
- **Supervised keyphrase extraction** = binary classification of keyphrase candidates
- State-of-the-art but requires **training data**
- Keyphrase annotation is **highly subjective**
- **Noisy and unbalanced** training data for supervised keyphrase extraction

We create 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.

Disagreement on Keyphrases



Keyphrase Extraction as Positive Unlabeled Learning



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**

Contributions

- Two datasets with **multiple annotations**
- Treat non-selected phrases as **unlabeled rather than negative**
- Reweigh keyphrases based on first classifier prediction, **normalize and coreference**
- Future work: **Better evaluation measures for keyphrases**

lusterck@intec.ugent.be

@lusterck

lusterck.github.io

emnlp2016