



Supervised Keyphrase Extraction as Positive Unlabeled Learning

Lucas Sterckx†, Cornelia Caragea*, Thomas Demeester†, Chris Develder† *Ghent University - imec, *University of North Texas her clear on 19 sept has - this connects to ner ...

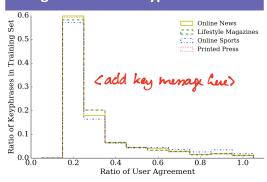
Our consibution:

Training Data for Supervised Keyphrase Extraction

- Supervised keyphrase extraction = binary classification of keyphrase Challeys/operismes/ problem? 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

maybe kerto add a key runage/condusia from His part?

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

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 target,Comet 67P/Churyumov-

I supped that sniph

another setting might be

test highlythy the object would be nia)

dear (but still just 1 lim of

Single Annotator

- Comet 67P/Churyumov-Gerasimenko
- **European Space Agency**
- Journey of four billion miles
- Six-minute thruster

Multiple Annotators

- Comet 67P/Churvumov-Gerasimenko
- **European Space Agency**
- Journey of four billion miles
- Six-minute thruster
- at 5 a.m. Eastern time

Positive Unlabeled Learning

Elkan, Nolito (2008) S < state core idea here!> $\frac{1}{m} \left(\sum_{\langle x, s=1 \rangle} h(x,1) + \sum_{\langle x, s=0 \rangle} w(x) h(x,1) + (1-w(x)) h(x,0) \right)$

Normalize + Coreference

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

key minage here?] 0.380 be a list hour? Users per document

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

make bur connect width

0.45/0.55 Six-minute thruster

0.03/0.97 at 5 a.m. Eastern time

Experiments

but the positive unlabeled learning part without a

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	<u>.381</u>	<u>.426</u>	.303	. <u>327</u>	1	1	1	1	1	1
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	. <u>305</u>	.322	.245	.194	.275	.200	.402	.434

Contributions

coreference

also indent 2nd line ... •Two datasets with multiple annotations

- Treat non-selected phrases as unlabeled rather than negative
- Reweigh keyphrases based on first classifier prediction, normalize and
- Future work: Better evaluation measures for keyphrases

There are recommendation? - provide a bit more sorker ... ☐ lusterck@intec.ugent.be





