K.D STRIKES BACK

Episode II

from Keyphrases to Labelled Domains Using External Knowledge Sources

WHAT

- L-KD (*Labelled-KD*): tool for keyphrase clustering and labelling
  - Extension of KD: **http://dh.fbk.eu/technologies/kd**
  - Based on external linguistic and knowledge resources: i.e., WordNet Domains and ConceptNet 5
  - Works on English and Italian texts
  - Online demo: **http://dh.fbk.eu/technologies/l-kd**

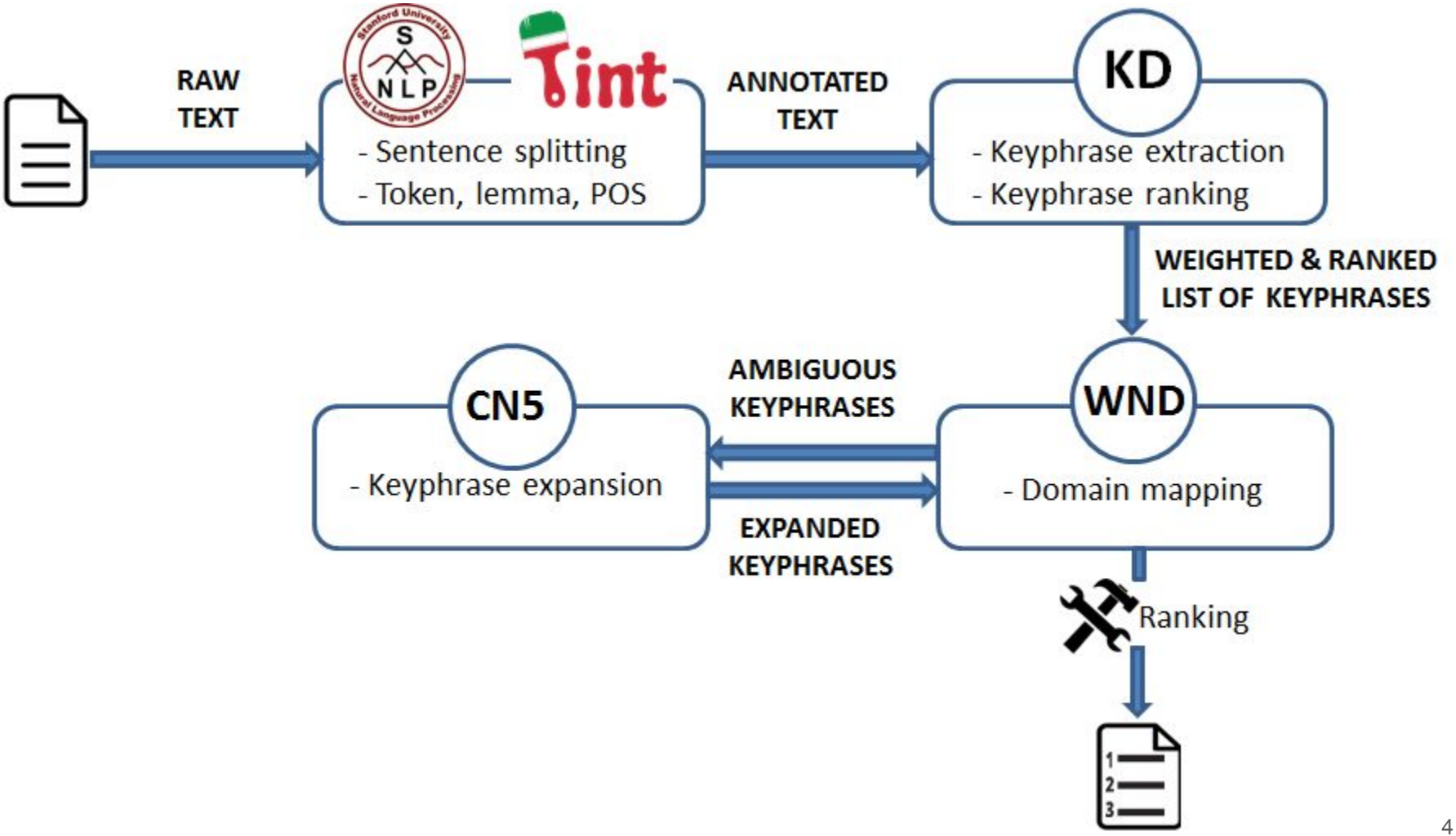
WHY

- Track the flow of information and retain only relevant content at two granularity levels: i.e., key-concepts and domains

- Simpler approach than topic modelling:
  - easier to be interpreted
  - based on a well-established domain hierarchy

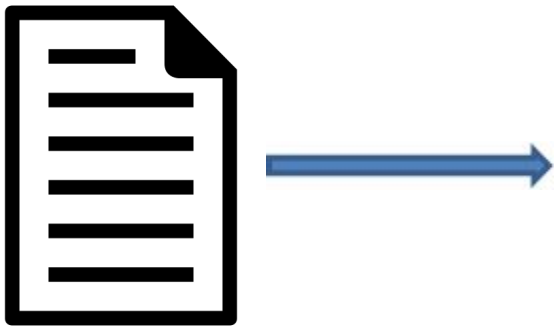
- Exploit a novel combination of WordNet Domains and ConceptNet 5

HOW


RAW TEXT
- Sentence splitting
- Token, lemma, POS
ANNOTATED TEXT
KD
- Keyphrase extraction
- Keyphrase ranking
WEIGHTED & RANKED LIST OF KEYPHRASES
WND
- Domain mapping
AMBIGUOUS KEYPHRASES
EXPANDED KEYPHRASES
CN5
- Keyphrase expansion
Ranking

HOW: STEP 1

- Text Pre-processing + Keyphrase extraction & ranking
  - Intermediate steps: sentence splitting, tokenization, lemmatization, part of speech tagging
- Output: list of single or multi-token keyphrases



KEYPHRASE	*FREQ*	*WEIGHT*
natural habitat	7	45.23425
ecological network	4	19.38611
species	6	19.38611
nature	3	9.693053

HOW: STEP 2

- Mapping of lemma forms of keyphrases with the lemmas in WordNet Domains (WND) aligned to WordNet 3.0
  - For Italian: Open Multilingual WordNet project
- Output: list of keyphrases associated to one or more domain

KEYPHRASE	*marsh*	*nature*
WND	*marsh 09347779 geography* **UNAMBIGUOUS**	*nature 09503682 Factotum* *nature 04623113 Psychological_Features* **AMBIGUOUS**

HOW: STEP 3

- Expansion of ambiguous keyphrases aligning them with lemmas in ConceptNet 5 (http://conceptnet5.media.mit.edu/) and exploiting hierarchical and synonymous relations

- Output: keyphrases extended with connected concepts

nature → *RelatedTo* → flora
nature → *RelatedTo* → environment
nature → *RelatedTo* → ecosystem
nature → *IsA* → great place
nature → *HasA* → many wonder
…..
…..

nature: flora, environment, fauna, ecosystem, great place, many wonder, country, conservation...

HOW: STEP 4

- Domain mapping of expanded keyphrases using WND (as in step 2)

- Output: list of domains associated to each expanded keyphrase

nature: flora, environment, fauna, ecosystem, great place, many wonder, country, conservation...

Biology = 19
Plants = 8
Animals = 5
…
...

HOW: STEP 5

- Creation of the final ranking
- Output: list of domains with associated keyphrases

Geography:	natural habitat
	river
	high water
	land
	marsh
Biology:	nature
	species

EVALUATION

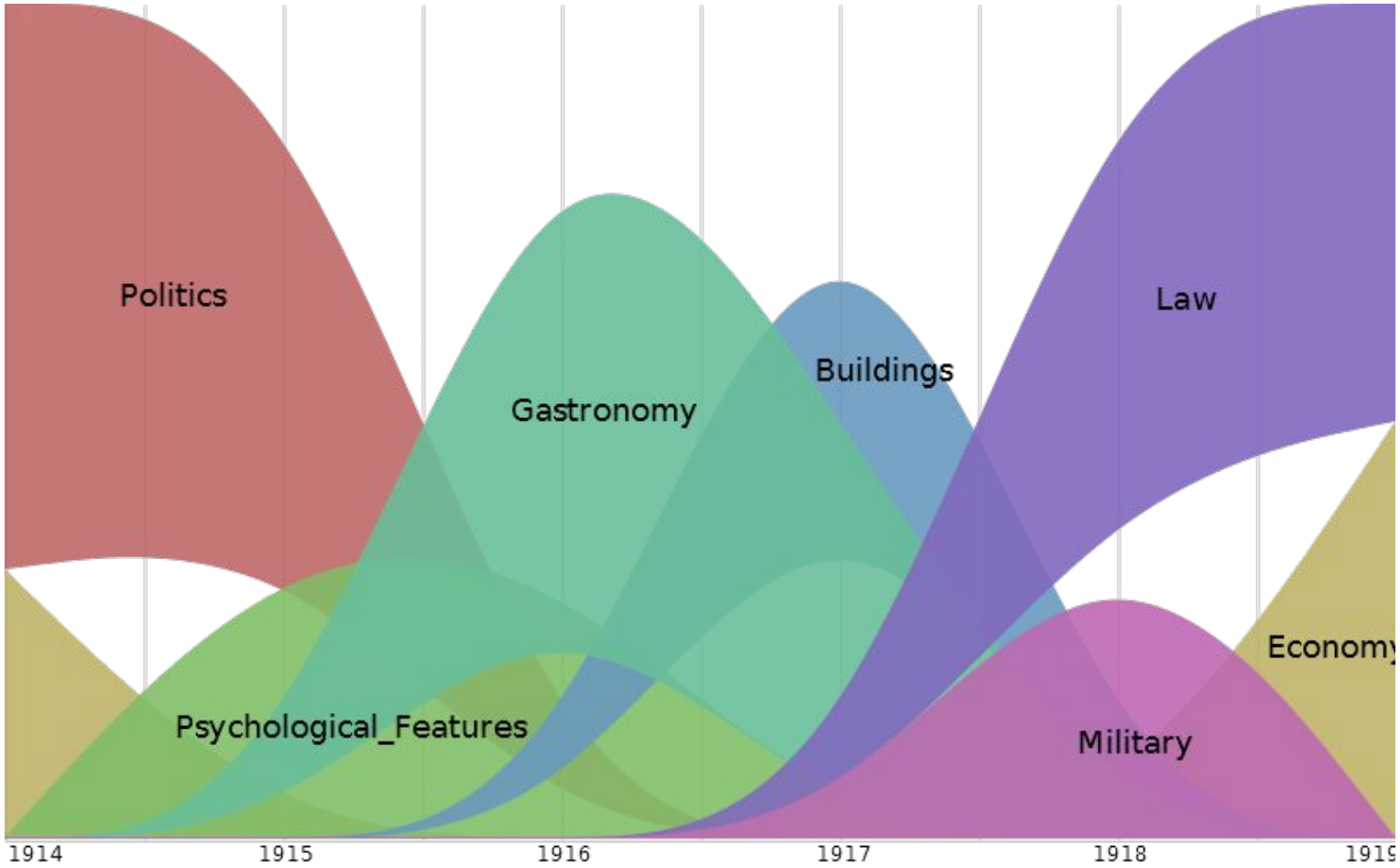
- 20 Newsgroup dataset
  - 20,000 documents manually assigned to one out of 20 different categories which in turn were mapped to domains
    - Categories: *rec.sport.baseball - rec.sport.hockey*
    - Domain: *Sport*
- 80% accuracy: perfect match between the first domain ranked by L-KD and the original category

rec.sport.baseball - rec.sport.hockey	**Sport**	game, playoff, second period
	Play	player, baseball

sci.electronics	Law	article, opinion, information
	Electricity	amateur radio, voltage, wire

USE CASE

- Alcide De Gasperi's writings



FUTURE WORKS

- Investigate open issues on Italian:
  - Find a suitable gold standard for the evaluation: use Wikipedia?
  - Extend the current mapping between Italian lemmas and WordNet 3.0
- Release L-KD as a standalone module

THANK YOU!

Rachele Sprugnoli
Giovanni Moretti
Sara Tonelli

Digital Humanities Group - FBK
http://dh.fbk.eu
@DH_FBK

DH DIGITAL HUMANITIES

FBK FONDAZIONE BRUNO KESSLER