Enhance Social Context Understanding with Semantic Chunks

Siqiang Wen, Zhixing Li, and Juanzi Li

Knowledge Engineering Group, Dept. of Computer Science and Technology, Tsinghua University

Outline

- Motivation & Challenges
- Related Work
- Approach
- Experiment
- Conclusion

Motivation

Wordssss!

PHRASES?

Motivation

... //@江南大野花: 哈哈哈哈哈这个我看懂了！不容易 //@夏目家的小诗哥: 哈哈哈还能再黑五百年//@疯人院小王子:滚啊//@蛋糕要蘸沙拉酱: 闭嘴！！哈哈哈哈好好看球！ //@_卖得赛_Pegasu:舌尖上的巴西

nsubj(这个-2, 哈哈哈哈哈-1)
root(ROOT-0, 这个-2)
nsubj(看懂-4, 我-3)
ccomp(这个-2, 看懂-4)
asp(看懂-4, 了-5)
neg(容易-8, 不-7)
rcmod(哈哈哈-9, 容易-8)
nsubj(闭嘴-18, 哈哈哈-9)
advmod(滚-16, 还-10)
mmod(滚-16, 能-11)
advmod(滚-16, 再-12)
amod(年-15, 黑-13)
nummod(年-15, 五百-14)
dep(滚-16, 年-15)
dep(闭嘴-18, 滚-16)
dep(滚-16, 啊-17)
conj(看懂-4, 闭嘴-18)
dep(这个-2, 哈哈哈哈-21)
advmod(看球-23, 好好-22)
vmod(舌尖-25, 看球-23)
lobj(上-26, 舌尖-25)
assmod(巴西-28, 上-26)
assm(上-26, 的-27)
dobj(哈哈哈哈-21, 巴西-28)

Challenges

- How to extract phrases without parser
- How to extract readable sematic dependency phrases
- The limit of labeled semantic dependency corpora

Related Work-supervised

- Improved Automatic Keyword Extraction Given More Linguistic Knowledge
  - Hulth Anette
  - In EMNLP '03
- Learning Algorithms for Keyphrase Extraction
  - Peter D. Turney
  - In Inf. Retr.
- KEA: Practical Automatic Keyphrase Extraction
  - Witten Ian H. and Paynter Gordon W. and Frank Eibe and Gutwin Carl and Nevill-Manning Craig G.
  - In DL '99

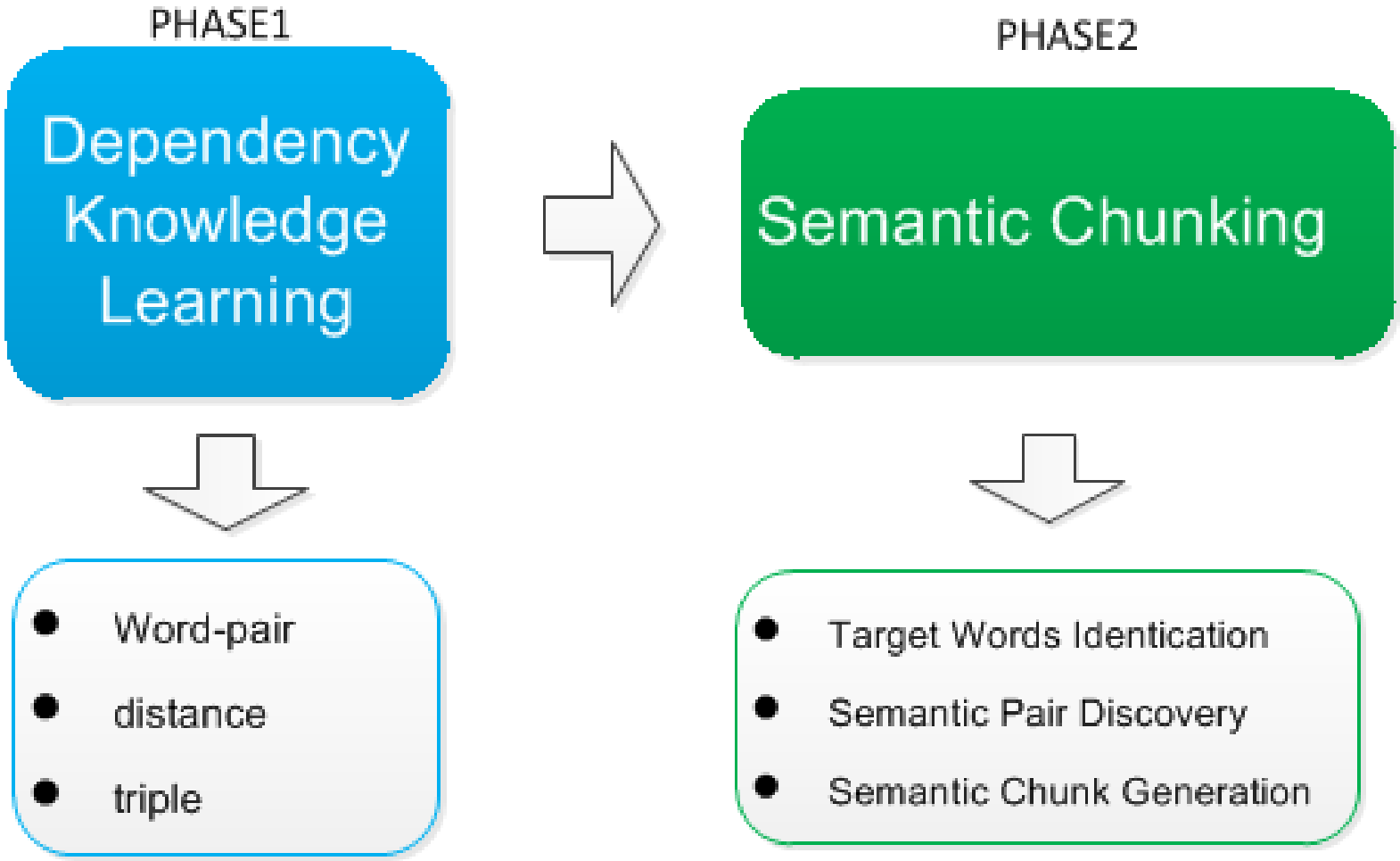
Related Work-unsupervised

- TextRank: Bringing Order into Text
  - Rada Mihalcea and Paul Tarau
  - In EMNLP '04
- Single Document Keyphrase Extraction Using Neighborhood Knowledge
  - Xiaojun Wan and Jianguo Xiao
  - In AAAI ’08
- Automatic Keyphrase Extraction via Topic Decomposition
  - Liu Zhiyuan and Huang Wenyi and Zheng Yabin and Sun Maosong
  - In EMNLP ‘10

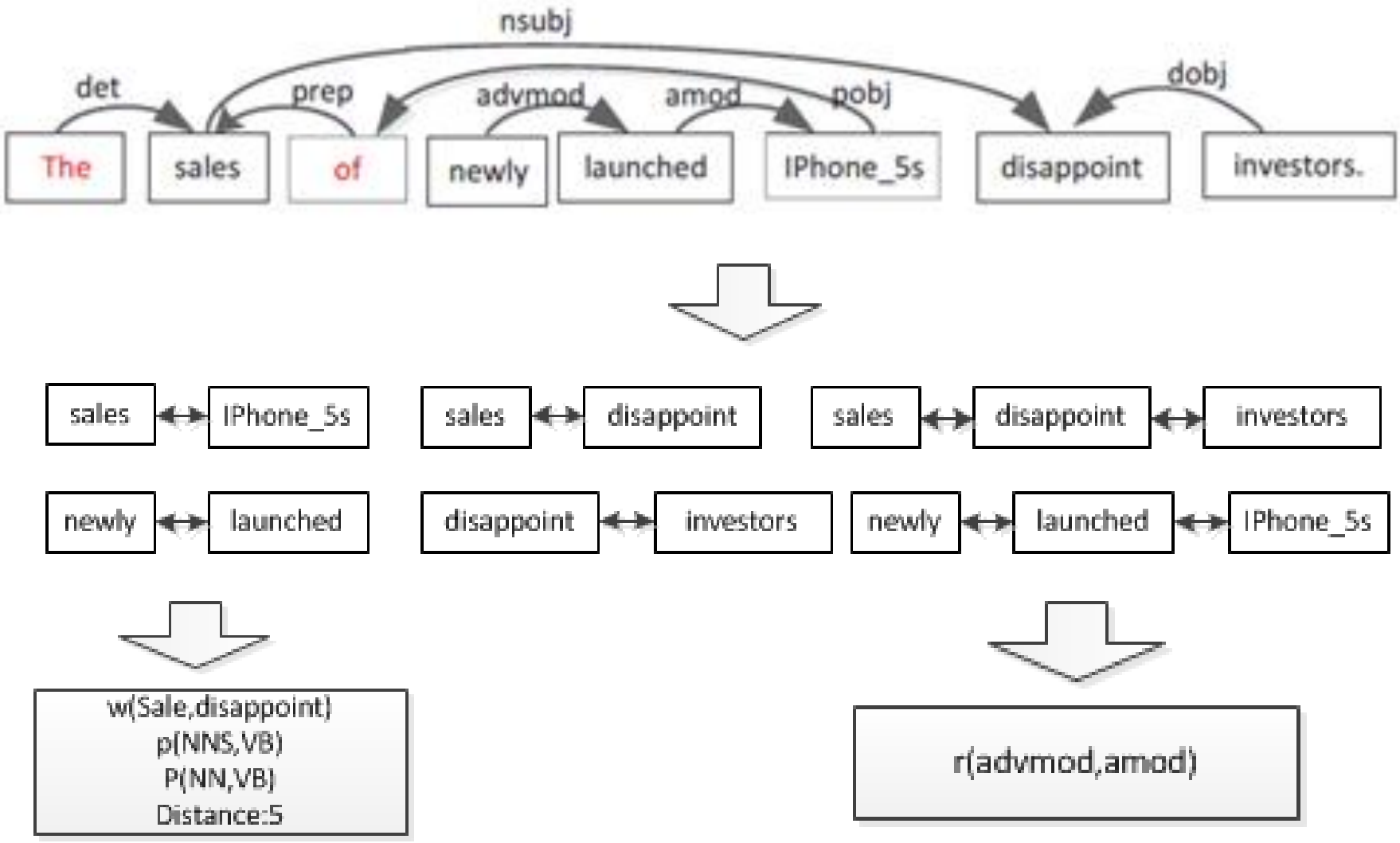
Related Work-social context

- Interest mining from user tweets
  - Vu Thuy and Perez Victor
  - In CIKM '13
- NE-Rank: A Novel Graph-Based Keyphrase Extraction in Twitter
  - Bellaachia Abdelghani and Al-Dhelaan Mohammed
  - In WI-IAT '12
- Topical Keyphrase Extraction from Twitter
  - Zhao Wayne Xin and Jiang Jing
  - In HLT '11

Approach Framework



Approach-Dependency Knowledge Learning



Approach-Semantic Chunking

- Semantic chunk
  - A phrase which is meaningful and significant expression describing the fist of given texts.

$$s_{d_i} = \begin{cases} (w_i), & w_i \in T_d \\ (w_i, w_j), \exists r_a (w_i, w_j, r_a) \in Kno_w \\ (w_i, w_j, w_k), \exists r_a (w_i, w_j, r_a) \in Kno_w, \\ \quad \exists r_b (w_j, w_k, r_b) \in Kno_w, \\ \quad (r_a, r_b) \in Kno_r \end{cases}$$

Approach-Semantic Chunking

- Target Words Identication
  - Target words can evoke a semantic phrase in a given sentence represents the dominant concepts in the social content
  - Noun, especially entity
- Semantic Pair Discovery
  - Tongyici Cilin for Chinese and WordNet for English

$$R(w_i, w_j) = \alpha R_w(w_i, w_j) + (1-\alpha)(|R_d(w_i, w_j) - d_{ij}|)^{-1}$$

$$R_w(w_i, w_j) = \begin{cases} C_w(w_i, w_j), & w_i, w_j \in W_K \\ \sum_{k=1}^{e}\sum_{l=1}^{e} p_{kl} S(w_{ik}, w_{jl}), & else \end{cases}$$

Extended by knowledge base

$$S(w_{ik}, w_{jl}) = C(w_{ik}, w_{jk}) \times Sim_{ik} \times Sim_{jl}$$

$$p_{kl} = \frac{\#N(w_{ik}, w_{jk})}{\sum_{k=1}^{e}\sum_{l=1}^{e} \#N(w_{ik}, w_{jk})}$$

Approach-Semantic Chunking

- Semantic Chunk Generation
  - Find Triples

$$C(w_i, w_j, w_k) = p(r_m|(w_i, w_j))p(r_n|(w_i, w_j))h(r_m, r_n)$$

  - select top-n of $C(w_i, w_j, w_k)$
  - (w_i, w_j, w_k) replaces (w_i, w_j) and (w_j, w_k)
  - Rank pairs and triples

Experiment

- Datasets
  - Sources(Chines:Weibo, English:Yahoo! News comments)

Table 3. Statistics on DataSets

Dataset	$\mid D \mid$	$\mid W \mid$	$\mid V \mid$	$\mid N_s \mid$	$\mid N_w \mid$
Sina(cn)	1000	129304	15318	7.06	18.32
Yahoo!(en)	1000	97392	10821	5.96	16.34

- N_s:the average number of sentences in a document
- N_w:the average number of words in a sentence

Experiment

- Evaluation Methods
  - Manually annotation
    - 3 annotators
  - Rules:
    - 3: Very good phrase, completely capturing gist of the document
    - 2: Reasonable and readable phrase, but not completely capturing gist
    - 1: Phrase is related to the contexts, but not readable
    - 0: Phrase is completely inappropriate

Experiment

- Baseline
  - Unsupervised Keyphrase Extraction Method:TextRank
  - Dependency Parsing

- Result

Table 4. Overall results of various methods for social contexts

	SmanC	*SmanC-KB*	*parser*	*textRank*
Sina(cn)	**2.237**	2.156	1.918	1.424
Yahoo!(en)	**1.871**	1.859	1.548	1.213

Conclusion

- We utilize part of semantic dependency relations between importance words learned from semantic dependency corpora and knowledge base to extract semantic chunks to capture the gist of the given document. The method does not need all relations between words like parser.
- We learn word knowledge, relation knowledge and distance knowledge from semantic dependency corpora. With these knowledge, we can extract long distance dependency phrases to form semantic chunks.

Thanks!