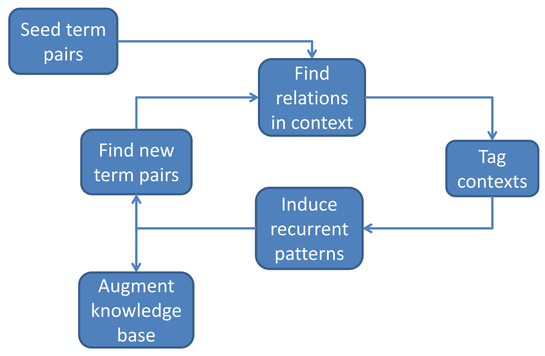
Extracting biomedical knowledge from the web using knowledge patterns

Jakob Halskov, Ph.D. (jakob_halskov@ubic.co.jp)
Senior Technical Researcher, Research Team, Behavior Informatics Laboratories
UBIC Inc., Meisan Takahama building 7F, 2-12-23 Kounan, Minato-ku, Tokyo 108-0075, Japan
http://www.ubic.co.jp

Abstract

With the emergence of Big Data **automatic knowledge extraction** has become an increasingly important issue, see for example the recent workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction at NAACL-HLT 2012. This poster describes the implementation and evaluation of a web-based biomedical knowledge mining system capable of detecting important **semantic relations between biomedical concepts** in unstructured data (free text). For example, the system can detect possible side-effects given a particular drug, possible drugs causing a particular effect, instances (hyponyms) of a particular drug and so forth.

1. Discovering Knowledge Patterns (KP) and semantic relations iteratively



Relation type (UMLS)	Seed term pair example ($term_1$ <KP> $term_2$)	Chinese equivalent
Induces	Aspirin <> bleeding	阿司匹林<>出血
May_prevent	Selenium<> DNA damage	硒<>DNA损伤
ISA	Haloperidol<>antipsychotic agent	氟哌啶醇<>抗精神病药
Synonymy	apoptosis<>programmed cell death	细胞凋亡<>程序性细胞死亡

2. Evaluating Knowledge Pattern precision

$$prec(KP) = \frac{\sum_{n=1}^{4} t_{1,n}; t_{2,n} \in R_{pos} \frac{C_{web}(t_{1,n}, KP, t_{2,n})}{C_{web}(t_{1,n}, *, t_{2,n})}}{\sum_{n=1}^{4} t_{1,n}; t_{2,n} \in R_{pos} \frac{C_{web}(t_{1,n}, KP, t_{2,n})}{C_{web}(t_{1,n}, *, t_{2,n})} + \sum_{n=1}^{4} t_{1,n}; t_{2,n} \in R_{neg} \frac{C_{web}(t_{1,n}, KP, t_{2,n})}{C_{web}(t_{1,n}, *, t_{2,n})}}$$

Where C_{web} is the web search engine hit count for the <term,KP,term> triplets, R_{pos} is a set of four positive term pairs for the target relation, and R_{neg} is a set of four negative term pairs for the same relation.

Synonymy KPs (examples)	ISA KPs (precision) [hypernym-hyponym]	ISA KPs (precision) [hyponym-hypernym]
or	e.g. (100%)	is an (100%)
see	such as (99.9%)	and other (99%)
also known as	including (99.2%)	is a new (79.8%)
Ie	like (89.6%)	or other (69.5%)
means	i.e. (77.2%)	see (59.2%)
also called	include (69%)	…

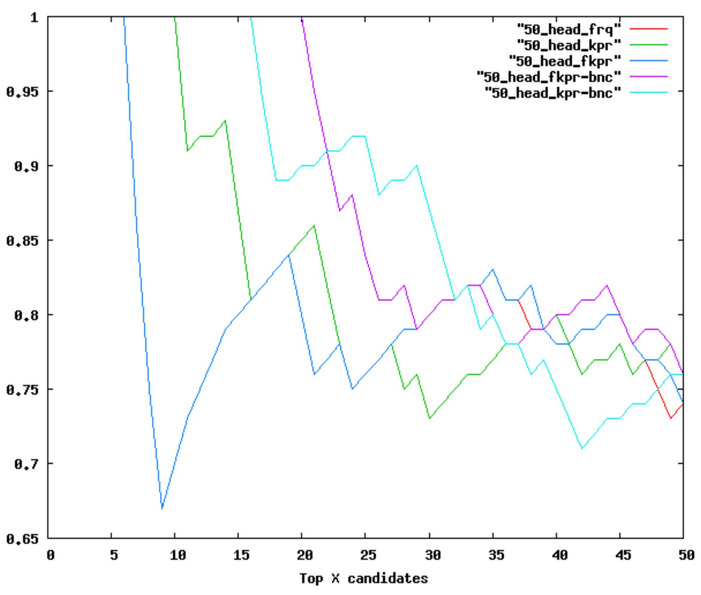
3. Defining & evaluating ranking schemes

$$frq(NP) = \sum_{KP \in P} C_{sample}(t, KP, NP)$$

$$kpr(NP) = |\{KP \in P | \exists (t, KP, NP)\}| \qquad kpr_{bnc}(NP) = \frac{kpr(NP)}{\log(C_{BNC}(NP_{head}))}$$

Scheme	Extracted noun phrases (NP) are ranked by the …
Head-frq	simple frequency of the head noun in the corpus of text snippets from the web
Head-kpr	range of different knowledge patterns co-occurring with the head noun
Head-fkpr	product of head-frq and head-kpr
Head-kpr/fkpr-bnc	head-kpr/fkpr scheme but penalizing head nouns occurring frequently in a corpus of general language (the British National Corpus).

Precision of semantic relation extraction for ISA(Haloperidol;?)



4. Reimplementation for Chinese

Using the free FAROO web search API there was too little data for calculating KP precision exactly as for English (i.e. using the equation of section 2). Instead statistical association scores vs. a large corpus of language for general purposes (LGP) were used to extract candidate Chinese KPs.

In the experiment reported below the following procedure was adopted.

1. Automatic extraction of web text snippets containing selected cMESH side effects
2. Automatic NLP (segmentation & part-of-speech tagging) of the texts
3. Automatic identification of verbs co-occurring with the side effects
4. Calculation of the association strength (log-likelihood ratio) between each verb (i.e. candidate KP) in the corpus of medical text snippets vs. an LGP corpus.

Selected cMeSH side effects/symptoms

Chinese	English
便秘	constipation
低血糖症	hypoglycemia
溃疡	ulcers
耳鸣	tinnitus
哮喘	asthma
支气管收缩	bronchoconstriction
变态反应	allergic reaction
荨麻疹	urticaria
出血	bleeding
细胞凋亡	apoptosis

KP discovery using a comparison of word frequencies in **analysis** corpus vs. **reference** corpus.

$$log-likelihood(KP|corpus) = 2\sum_{ij} O_{ij} log \frac{O_{ij}}{E_{ij}}$$

Computing association strength of verbs with medical corpus, cf. Evert (2005)

	Observed frequencies (O)			Expected frequencies (E)	
	Corpus= medical snippets	Corpus≠ medical snippets			
Phrase=KP	O_{11}	O_{12}	$=R_1$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
Phrase≠KP	O_{21}	O_{22}	$=R_2$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$
	$=C_1$	$=C_2$	=N		

Candidate Chinese Knowledge Patterns (causal relations)

Keyness (log-likelihood ratio)	Frequency (reference corpus)	Candidate KP (verb preceding side effect)	Semantic relation	English translation
27.17	0	防治	may_prevent	prevent
10.86	0	诱发	induce	cause/induce
10.86	0	挥拳	n/a	attack
6.60	123	服用	(may_prevent)	take medicine
5.89	428	治疗	may_prevent	cure
…	…	…	…	…
2.77	84	抑制	may_prevent	inhibit
2.12	303	下降	may_prevent	decrease
1.95	139	防止	may_prevent	prevent
1.53	182	表明	n/a	indicate
…	…	…	…	…
-0.04	664	引起	induce	lead to
…	…	…	…	…
-0.26	1647	发生	(induce)	arise/occur

References

Brin, S. (1998). ”Extracting Patterns and Relations from the World Wide Web”. In: *Proceedings of WebDB Workshop of the Sixth International Conference on Extending Database Technology* (EDBT), pp. 172-183.

Evert, S. (2005). Zur statistischen Analyse von Wortkombinationen: Wortpaare und Kollokationen. PhD Thesis. Institut für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart

Halskov, J.; Barrière, C. (2008). ”Web-based etraction of semantic relation instances for terminology work”. In: *Probing Semantic Relations*, pp. 19-42. Amsterdam: John Benjamins.

Meyer, I. (2001). ”Extracting knowledge-rich contexts for terminography. In: *Recent Advances in Computational Terminology*, pp. 279-302. Amsterdam: John Benjamins.

Wang, X.; Thompson, P.; Ananiadou, S. (2012). ”Biomedical Chinese-English CLIR Using an Extended cMeSH Ressource to Expand Queries”. In: *Proceedings of the Eighth International Conference on Language Recources and Evaluation* (LREC 2012), pp. 1148-1155.