Learning the Species of Biomedical Named Entities from Annotated Corpora

Xinglong Wang and Claire Grover

LREC
29 May 2008

1. Background and Motivation

2. Tagging Species to Biomedical Named Entities
   - Datasets and Ontologies
   - Detecting the Species Words
   - Rule-based Species Tagging
   - Machine-learning based Species Tagging

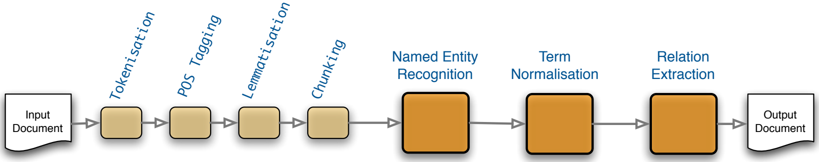
3. Conclusions and Future Work

Text Mining from Biomedical Literature

- Document Selection - Text Classification
- NLP Pipeline
  - NER - Named-entity recognition, Proteins, Tissue, Cellline, etc
  - TI - Term Identification (i.e., Normalisation) - Proteins, Genes, Tissue, etc
  - RE - Relation Extraction - Protein-protein interactions, Tissue Expression, Parent-Fragment

Text Mining from Biomedical Literature

The TXM text mining pipeline:



Example

Rrs1p has a two-hybrid interaction with L5.

Example

Rrs1p has a two-hybrid interaction with L5.

- Two proteins of species Saccharomyces cerevisiae (4932) normalised to the RefSeq identifiers NP_014937 and NP_015194

Example

Rrs1p has a two-hybrid interaction with L5.

- Two proteins of species Saccharomyces cerevisiae (4932) normalised to the RefSeq identifiers NP_014937 and NP_015194
- One experimental method

Example

Rrs1p has a two-hybrid interaction with L5.

- Two proteins of species Saccharomyces cerevisiae (4932) normalised to the RefSeq identifiers NP_014937 and NP_015194
- One experimental method
- A direct, positive and proven relation between both proteins

Example

Rrs1p has a two-hybrid interaction with L5.

- Two proteins of species Saccharomyces cerevisiae (4932) normalised to the RefSeq identifiers NP_014937 and NP_015194
- One experimental method
- A direct, positive and proven relation between both proteins
- A relation attribute specifying that the interaction was detected using the experimental method

Term Identification

Term Identification (TI) System: a system that grounds a biological term to a specific identifier in a reference database. A TI system usually comprises of:

- Ontology processor
- Matching system
  - NER and Approximate search
  - Brute-force approximate search
- Disambiguator/Filter - **species disambiguation**

Term Identification (Continued)

Variations of synonyms to terms and ambiguity in species often cause difficulty to TI:

- hRXRα: $\{$*RXR*α; *retinoid X receptor*, *alpha*; *NR2B1*$\}$
- *RXR*α: $\{$NP_002948 (human), NP_035435 (mouse), etc.$\}$

- E.g., abbreviation/acronym and normalising sequential characters.
- Species indicating characters, e.g., 'h' in *hRXR*α.

Species Tagging

- Species is essential for TI.
  - Database identifiers are species specific (e.g., RefSeq and UniProt)!
  - Interacting proteins in the BioCreAtIvE II IPS dataset belong to over 60 species.
  - Biomedical entities in the TXM EPPI dataset belong to 112 species, and those in the TE dataset belong to 61 species.
- Species tagging improves TI.
  - Our previous work (Wang, 2007) shows that species tagging improved performance of a rule-based TI system by 10%.
  - Further evidence to come (Wang and Matthews, BioNLP 2008).

Datasets and Ontologies

- The TXM corpora (EPPI and TE): various types of entities manually recognised and normalised. (Alex et al. 2008)
- Entities are normalised to identifiers of various databases (e.g., RefSeq, EntrezGene, MeSH).
- They are also "species-normalised" to NCBI Taxonomy identifiers.

TaxID	Name	Rank
8353	Xenopus	genus
262014	Xenopus	subgenus
8364	Xenopus tropicalis	species

Table: Taxonomy records for Xenopus in the NCBI taxonomy. 'Rank' refers to the hierarchy level of the node in the ontology.

Detecting the Species Words

1. .. expressed the endogenous mouse REST (mREST) ...
2. The sequences of the human and mouse CDK12S ...
3. .. CYP2B6, a human relative of CYP2B10 ...
4. The Drosophila methyl-DNA binding protein MBD2/3 ...

Detecting the Species Words (Continued)

- A lexical look-up component.
- Detecting words indicating species by searching 4 lexicons using rules written in *lxtransduce* grammar.
- The lexicons were derived from the NCBI Taxonomy and UniProt.
- They also contain hand-compiled Latin and English forms for a number of frequent species and allow for pluralisation (e.g., *mice*), adjectives (e.g., *ovine*) and different tokenisations (e.g., *E. coli*).

Species Tagging using the Species Words

Identify the species of a biomedical entity by looking at the nearby species words, using 4 simple rules:

1. *PrevWd*: assign the entity the species indicated by its preceding species word (if there is any).
2. *PrevWd Spread*: spread the species to all the entities with the same surface form in the article.
3. *PrevWd in Sent*: assign the entity the species indicated by the species word in the same sentence.
4. *PrevWd in Sent Spread*: spread the species to all the entities with the same surface form in the article.

Results

	PrevWd			PrevWd in Sent		
	P	R	F1	P	R	F1
EPPI	81.9	1.9	3.7	60.8	5.2	9.5
TE	91.5	1.6	3.2	56.2	7.8	13.6

	PrevWd Spread			PrevWd in Sent Spread		
	P	R	F1	P	R	F1
EPPI	63.9	14.2	23.2	39.7	50.5	44.5
TE	77.8	18.0	29.2	31.7	46.7	37.4

Table: Results (%) of the rule-based species tagger.

Revisiting the Examples

1. .. expressed the endogenous mouse REST (mREST) ...
2. The sequences of the human and mouse CDK12S ...
3. .. CYP2B6, a human relative of CYP2B10 ...
4. The Drosophila methyl-DNA binding protein MBD2/3 ...

For the last example:

TaxID	Name	Rank
7215	Drosophila	genus
7227	Drosophila melanogaster	species

Machine-learning Based Tagging

- Learn a probabilistic model using the training data that can predict species of an entity based on its surrounding context.
- Maximum entropy modeling was used (Software tool *maxent* was developed by Le Zhang at Edinburgh University).
- Features included **contexual words**, **previous nouns**, **previous adjectives**, **nearby species words**, and all **species** that occur in the document in question (as indicated by the species words).

Results

	BL	EPPI Model	TE Model	Combined Model
EPPI	60.56	**73.04**	66.42	71.28
TE	33.28	63.91	**70.73**	68.18
avg.	46.92	68.12	68.62	**69.73**

Table: Accuracy (%) of the machine-learning based species tagger tested on EPPI and TE *devtest* datasets. BL denotes the majority baseline, EPPI model was trained on the EPPI training dataset, TE model trained on the TE training dataset, and Combined model trained on a joint dataset of EPPI and TE.

Conclusions

- Rules relying on the “species words” can achieve high precision (81.88% and 91.49% on EPPI and TE) but very low recall.
- Spreading the species helped a little but not satisfactory.
- A maximum entropy classifier with a large set of selected features achieved F1 scores of 71.28% and 68.18% on EPPI and TE.
- However, the distributions of the species in the training data tend to bias the machine-learned models.

Future Work

- Measuring the impact of species tagging to term identification. (See Wang and Matthews, BioNLP 2008)
- Measuring the impact of species tagging to relation extraction.
- Using rules as constraints in machine learning based species tagging.

Thank you!

Thank you!

- The TXM project was funded by ITI Life Sciences, Scotland.
- The TXM Team: Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang