



Report on BioCreative Task II

Frédéric Ehrlér and Patrick Ruch

University of Geneva, Geneva
Swiss Federal Institute of Technology, Lausanne
Frederic.Ehrlér@cui.unige.ch
Patrick.Ruch@epfl.ch

29/03/2004

1

Plan



- Introduction
- Data provided
- Task 2.1
 - Methode
 - Result
 - future
- Task 2.2
 - Methode
 - Result
 - future



29/03/2004

2



Passage Retrieval and Text Categorization

- 2.1 'Recover' text that proofs the GO annotation:
 - Protein, and GO annotation provided for the publication
 - Participants will have to provide a part of the document that would (to a human expert) prove the original annotation
- 2.2 Provide GO annotation for human proteins:
 - Protein provided for the publication
 - Participants will have to
 - 'annotate' automatically the protein according to the information in this paper
 - provide a part of the document to prove the annotation

29/03/2004

3




Data provided

- 622 training data (for GO annotation)
- Gene ontology
 - Content:
 - Controlled vocabulary (ontology)
 - 3-Axes Ontology
 - Molecular Function Ontology
 - Biological Process Ontology
 - Cellular Component Ontology

29/03/2004

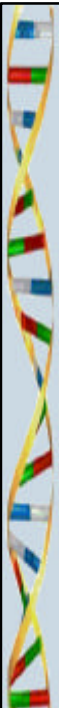
4



Data provided: Gene Ontology

- Different type of relations.
 - “Is a” relations
 - “Part of” relations
- 16648 Concepts
 - Between 1 and 8 term in a concept
 - Some have synonyms (same concept described by different words)

29/03/2004 5



Task 2.1

- The search has been done only for the abstract and not for the full article.
- As we haven't any training data it was difficult to train a tool.
- Only based on GO evidences ! (no use of the protein)

29/03/2004 6



Task 2.1

- Goal: Find the part of the document that motivates the annotation:
- We need:
 - Rule-based sentence splitting:
 - First identify and replace by markers some structures
 - Acronym: DNA.
 - Name: J. H. Smith
 - Number: 85.5
 - Use regular expression to split the sentence if we find the pattern «. »

29/03/2004

7



Methods

- 2 matching methods to score sentence are used:
 - Exact matching:
 - Count the number of «direct» match between the terms of the category and the sentence
 - Fuzzy matching: String edit distance
 - The second used is a fuzzy metric that computes a relation of similarity between two strings (Smith Waterman Distance)
 - The similarity between strings is the value of the alignment between two strings that maximizes the total alignment value

29/03/2004

8



Fuzzy matching

- Based on the Edit Distance
 - Three basic operations with the same cost:
 - The number of substitution
 - The number of insertion
 - The number of deletion

Excused	Source string	0
Exhused	Substitute "h" for "c"	1
Exhauled	Insert "a"	2
Exhausted	Insert "t"	3
Exhausted	Target string	3

29/03/2004

9



Task 2.1

- Combination of methods
 - Low discriminative power of exact match:
 - High precision or no result
 - Good discriminative power of fuzzy method
 - Low precision but good recall
 - Exact metric is more effective than fuzzy one but can't well discriminate the sentence, but 70% percent of the match occur without exact match
 - ✍ Combination of both to score the sentences

29/03/2004

10




Task 2.1: Example

- Searched Terms: protein serine/threonine kinase activity
- Sentences:
 - 1) Cdc42-induced **activation** of the mixed-lineage **kinase** SPRK in vivo.
 - 2) Src homology 3 domain (SH3)-containing proline-rich protein **kinase** (SPRK)/mixed-lineage **kinase** (MLK)-3 is a **serine/threonine kinase** that upon overexpression in mammalian cells **activates** the c-Jun NH(2)-terminal **kinase** pathway.
 - 3) This is, to the best of our knowledge, the first demonstrated example of a Cdc42-mediated change in the in vivo phosphorylation of a **protein kinase**.

	Direct Match	Smith Waterman	Levenshtein	Jaccard	Jaro	Final Score
1)	1	19	-45	0.062	0.62	29
2)	2	51	-18	0.12	0.58	71
3)	2	18	-12	0.083	0.58	38

29/03/2004

11

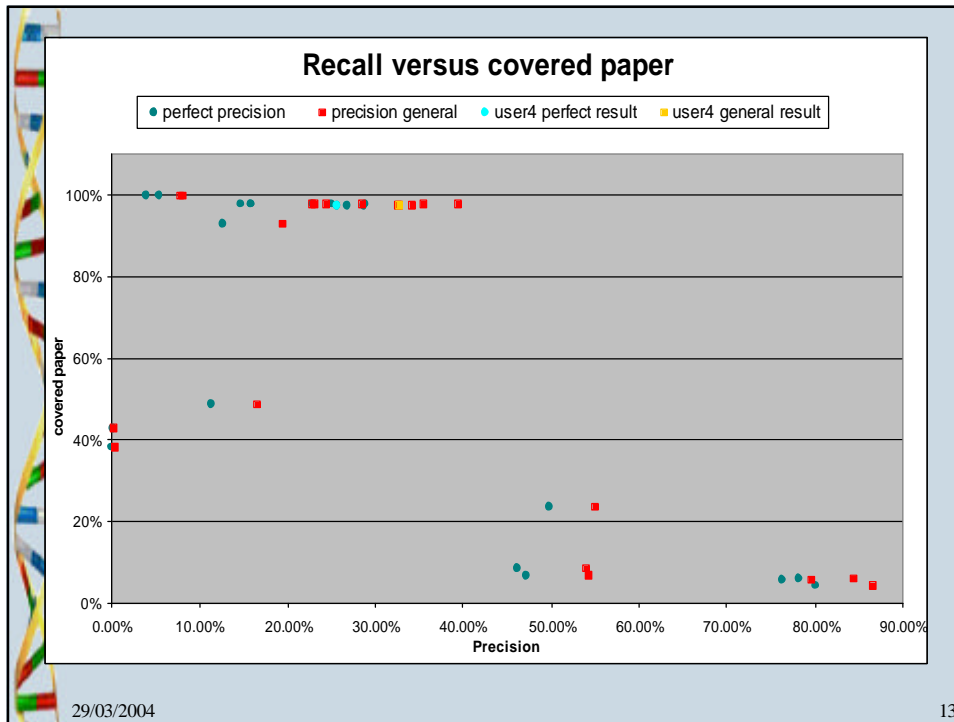


Task 2.1 : Results


- Number of test data: 1076
- Number of sentence evaluated: 1048
 - Perfect sentence: 268 (25.57%)
 - correct protein, general GO: 74 (7.06%)
- Simple methode but behave correctly

29/03/2004

12



Task 2.1: Future work



- Use
 - Hierarchy
 - only 30% of document have perfect match with a GO term
 - Full text article
 - More elaborated string normalization
 - GO definitions (features expansion)

29/03/2004 14





Task 2.2

- Usual task for automatic text categorization
- Categorize then apply the same tools as task 2.1 to retrieve the relevant sentence
- For each document several terms are asked

29/03/2004

15



Task 2.2 Data

- Work on the abstract
- Some figures
 - 660 categories
 - 725 features
 - 622 documents
 - 1869 relations between documents and categories

29/03/2004

16



ATC General Strategies

- *Empirical learning of text-concept associations* from a training set of texts and their associated concepts:
 - Reuters (Bayesian classifiers, Lewis 1992): 100 classes
 - Text categorization/filtering paradigm [Sebastiani: hundreds...]

✍ Effective but Learning Conditions...
- *Retrieval based on word-matching*, which attributes concepts to text based on shared words between the text and the concepts:
 - Rare: cf. SAPHIRE Int., Hersh et al. 1998

✍ **Recycling a MeSH categorization tool !**

29/03/2004

17



Fusion of Classifiers

- FSA Pattern matcher + thesaurus [RegEx]
 $word_1 \dots word_n \not\Leftarrow word_1 \dots _ [* , 2] \dots word_n$
 $word_1 \dots word_n \not\Leftarrow word_1 \dots [word_i]^* \dots word_n$

✍ Boolean scoring
- Vector Space: Porter stems + TF*IDF weighting [VS]

✍ Cosine distance/Similarity
- + Noun Phrase Indexing
- One classifier per GO axe: funct/comp/proc

29/03/2004

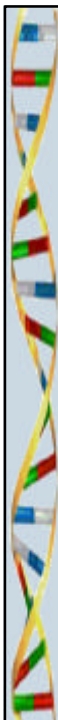
18



Vector space parameters

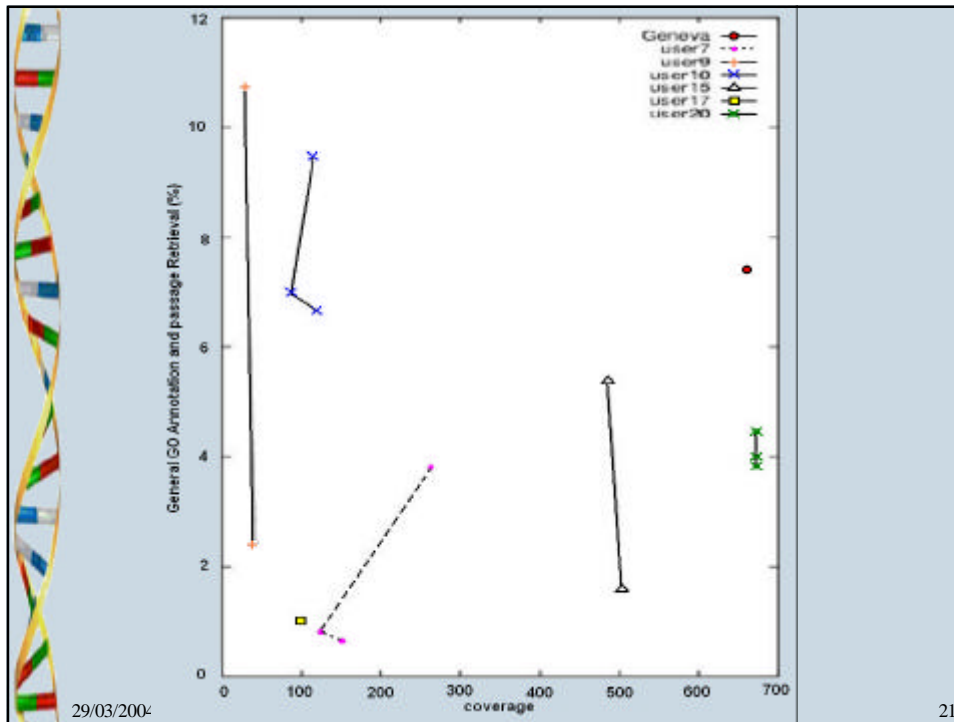
- TF: $\text{weight}_{\text{term}} = f(\text{term frequency})$
- IDF: $\text{weight}_{\text{term}} = f(\text{document frequency}^{-1})$
- Normalization: to balance long and short documents
- Maximizing mean average precision on a MeSH mapping task !

Term Frequency	
First Letter	$f(tf)$
n (natural)	tf
l (logarithmic)	$1 + \log(tf)$
a (augmented)	$\alpha + \beta \times (\frac{tf}{\max(tf)})$, where $\alpha = 1 - \beta$ and $0 < \alpha < 1$
Inverse Document Frequency	
Second Letter	$f(\frac{1}{df})$
n(no)	1
t(full)	$\log(\frac{N}{df})$
Normalization	
Third Letter	$f(\text{length})$
n(no)	1
c(cosine)	$\frac{1}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{i=1}^t w_{i,q}^2}}$



Results 2.2: User 4

- Number of sentence evaluated: 661
 - Perfect sentence: 78 (11.8%)
 - correct protein, general GO: 49 (7.41%)




21

Future Work

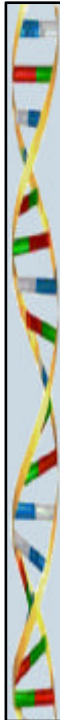
- Tune the classifier for the GO ontology
- Use GO definition
- Use the hierarchy
- Use the full-text article
- Use argumentative moves:
avoid « Materials and Methods » sections !

... Test it on the next TREC Genomic Track



29/03/2004

22



Thank you for your attention !

Acknowledgement:

- Christian “Responsive” Blaschke
- All the BioCreative team
- Christine Chichester (GeneBio SA)

Download:

<http://lithwww.epfl.ch/~ruch/softs/softs.html>

29/03/2004

23