Learning of Semantic Relations between Ontology Concepts using Statistical Techniques

A. Tegos[1,2], V. Karkaletsis[1], A. Potamianos[2]
tegos@iit.demokritos.gr, vangelis@iit.demokritos.gr,
potam@telecom.tuc.gr

[1]Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece
[2]Department of Electronics and Computer Engineering, Technical University of Crete, Greece

High-level Information Extraction Workshop 2008 (HLIE08), ECML-PKDD 2008

Introduction

- A methodology for automatic learning of ontologies from texts which are semantically annotated with instances of ontologies' concepts
- Applying statistical techniques to metadata extracted from the annotated texts we discover:
  - semantic relations among the annotated concepts
  - cardinality restrictions for these relations
- The method was applied to corpora from two different domains, *athletics* and *biomedical*, and was evaluated against the existing manually created ontologies for these domains

Outline

Introduction

The Proposed Method

- Finding the Semantic Relations of concepts
- Finding the Cardinality Restrictions

Experimental Assessment

Conclusions

Future Plans

Basic assumption

Our method is based on the assumption that concepts which are semantically related, tend to be "near" as context in a plain text

- This assumption arises from the principle of coherence on linguistics

The discovery process is not based to commonly used assumptions:

- Verbs typically indicate semantic relations
- Does not exploit lexico-syntactic patterns or clustering methods
- Does not use any external knowledge sources like WorldNet

Definitions

- **Low-Level**: concepts whose instances are associated with relevant text portions
  e.g. *name(has-instance)* or the *age(has-instance)*

- **High-Level**: “compound” concepts in such a way that instances of these concepts are related to instances of low-level concepts
  e.g. *person(name, age, nationality, gender)*

- We focus on the discovery of semantic relations between high-level concepts, but we also show the applicability of the proposed approach to low-level concepts

Requirements

- The method requires the annotation of the corpus with instances of ontology's concepts.
- In the case of high-level concepts as instances we consider the fillers of the concept's attributes that have been found in a document.

An example of the annotation

The 34-year-old, World marathon record holder and two-time Olympic and four-time World 10,000m champion Haile Gebreselassie of Ethiopia today announced that he intends to compete in this 2008 FKB-Games - IAAF World Athletics Tour - in Hengelo, the Netherlands on 24 May in his bid to make Ethiopia's team for the Beijing Olympics in China.

Athlete (name:*Haile Gebreselassie*, age:*34*, nationality: *Ethiopia*, gender:*NotFound*)
SportsCompetition (sport-name:*10,000m*, city:*Hengelo*, stadium-name: *NotFound*, date:*24 May*)

The proposed method

The proposed method for ontology learning involves 2 major steps:

- Finding the semantic relations of concepts that have been annotated in the corpus.
- Finding the cardinality restrictions for the extracted relations.

1. Finding the offsets of the annotated instances

- Based on our assumption, we treat each document of the corpus as a sequence of symbols.
- In this manner, each document is represented in a one-dimensional Euclidean space, depending on the place in which each symbol is found in the text.
- We find for each document the offsets of the annotated instances.
- As offset of an instance is defined the set that represents the minimum part of text which encloses all its fillers.

Example for the offset of the annotated instances

The 34-year-old, World marathon record holder and two-time Olympic and four-time World 10,000m champion Haile Gebreselassie of Ethiopia today announced that he intends to compete in this 2008 FKB-Games - IAAF World Athletics Tour - in Hengelo, the Netherlands on 24 May in his bid to make Ethiopia's team for the Beijing Olympics in China.

Athlete (name:*Haile Gebreselassie*, age:*34*, nationality: *Ethiopia*, gender:*NotFound*)

SportsCompetition (sport-name:*10,000m*, city:*Hengelo*, stadium-name: *NotFound*, date:*24 May*)

- The offset of the document is the set $[0, 342]$.
- The offset of the phrase "*34-year-old, World marathon*" is the set $[4, 30]$
- The offset for the *Athlete*'s instance is the set $[4, 134]$.
- The offset for the *SportsCompetition*'s instance is the set $[87, 270]$

2. Finding overlapping instances

- For each document, we search for the different pairs of concepts that have overlapping instances:

For the document doc_z*, of the corpus:*
$C_{doc_z} = \{C_1, C_2, \ldots, C_n\}$ *where* $C_i = \{I_1, I_2, \ldots, I_m\}$
where $I_k = [l, r] \bigcap \mathbb{N}$ *and* $l < r$,
we compare the instances' offsets:
$\forall (I_x, I_y)$ *where* $I_x \in C_i$, $I_y \in C_j$
and $C_i \in C_{doc_z}$ *and* $C_j \in C_{doc_z} - \{C_i\}$

$$If \left(I_x \bigcap I_y \neq \emptyset \right) \quad then\ create\ a\ pair\ \left(C_i, C_j \right)\ for\ doc_z \qquad (1)$$

- Note that for each document we are interested only in finding the different pairs of related concepts and not the number of occurrences for each of these pairs.

3. The semantic-correlation metric

- This metric measures the tendency of concept C_i to be semantically related, either taxonomically or non-taxonomically, with concept C_j, but not the inverse.

$$S(C_i \rightarrow C_j) = P(C_j|C_i)\cdot \left(1 + I(C_i, C_j)\right) = \\ = P(C_j|C_i)\cdot \left(1 + log\left(\frac{P(C_j|C_i)}{P(C_i)\cdot P(C_j)}\right)\right) \qquad (2)$$

- This definition is based on our assumption that concepts which are semantically related, tend to co-occur “near”. Therefore, concepts whose instance offsets overlap frequently tend to be semantically related.

3. The semantic-correlation metric (cont'd)

- We use in our metric the conditional probability $P(C_j|C_i)$, in order to find for the concept C_i the most probable concept C_j with which is frequently overlapped.

- We use the mutual information in order to enhance our metric with the association between the concepts C_i and C_j.
  - strong association between C_i and C_j: $P(C_j|C_i) > P(C_i) \cdot P(C_j)$, $I(C_i, C_j) > 0$
  - no interesting association between C_i and C_j: $P(C_j|C_i) \approx P(C_i) \cdot P(C_j)$, $I(C_i, C_j) \approx 0$
  - if C_i and C_j are not associated: $P(C_j|C_i) < P(C_i) \cdot P(C_j)$, $I(C_i, C_j) < 0$

3. The semantic-correlation metric (cont'd)

- We compute the semantic-correlation scores between C_i and each of the rest of the concepts. The concept that maximizes this score is the concept with which the concept C_i is related to.

Find how concepts are related:
$C_{corpus} = \{C_1, C_2, \ldots, C_n\}, \quad \forall C_i \in C_{corpus},$

$$RELATE \quad C_i \rightarrow C_j, \quad \arg\max_{C_j} S\Big(C_i \rightarrow C_j\Big), \tag{3}$$

where $C_j \in C_{corpus} - \{C_i\}$

Discovery of semantic relations between low-level concepts

- We apply the proposed methodology with a variation on the definition of the instance offset of each low-level concept.

- We extend the offset of each instance by X symbols to the left and to the right.

- The usage of a window size, is motivated by the fact that instances of low-level concepts contain very few words and thus semantically related concepts might be near each other in the text but not overlapping.

Finding the cardinality restrictions for the discovered relations

- The types of connectivity that our methodology is able to specify, are $(1:N)$, $(N:1)$ and $(M:N)$
- The proposed methodology for the discovered relation $C_A \rightarrow C_B$ consists of the following steps:

1. For each document in the corpus that contains instances of the concepts $C_A = \{I_{A_i}, \dots\}$ and $C_B = \{I_{B_j}, \dots\}$, we create a list with the overlapping instances, of the concepts C_A and C_B.
2. For each list, we find the type of connectivity, for each document, between the instances of concepts C_A and C_B as follows:

$$\left.\begin{array}{l} I_{A_i}, I_{B_j} \\ I_{A_i}, I_{B_m} \\ \dots \end{array}\right\} \Rightarrow (1:N) \; or \; \left.\begin{array}{l} I_{A_i}, I_{B_j} \\ I_{A_k}, I_{B_j} \\ \dots \end{array}\right\} \Rightarrow (N:1) \; or \; \left.\begin{array}{l} I_{A_i}, I_{B_j} \\ I_{A_j}, I_{B_k} \\ \dots \end{array}\right\} \Rightarrow (M:N)$$

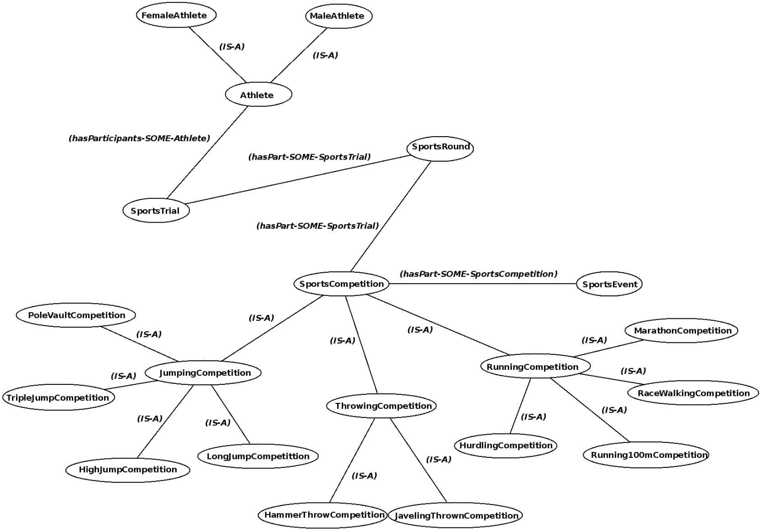
3. We specify as cardinality restriction, for the related instances of concepts C_A and C_B, the type of connectivity that occurs more often in the corpus.

Setting the Experiments

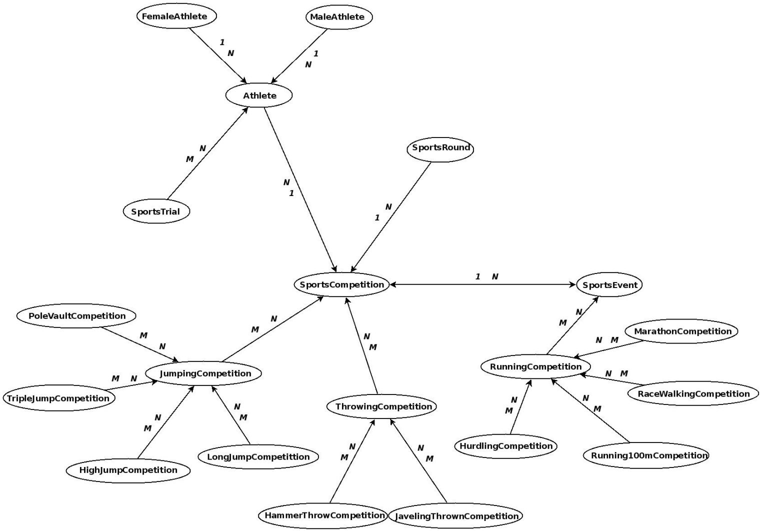
- The proposed method was applied on two corpora of different domains and the extracted ontologies were evaluated with respect to the corresponding manually created ontologies.

- The first corpus is on athletics domain, was obtained from BOEMIE project
  - 2,087 web pages containing athletic articles for 10 different sports competitions, mainly from IAAF web site
  - contains 36,240 instances' annotations, for 20 high-level concepts
- The second corpus is on biomedical domain
  - 286 abstracts of Pubmed
  - contains 1887 instances' annotations, for 6 high-level concepts

The manually created ontology for the domain of athletics



The automatically extracted ontology for the domain of athletics



The extracted and the manually created ontology for the domain of biomedical

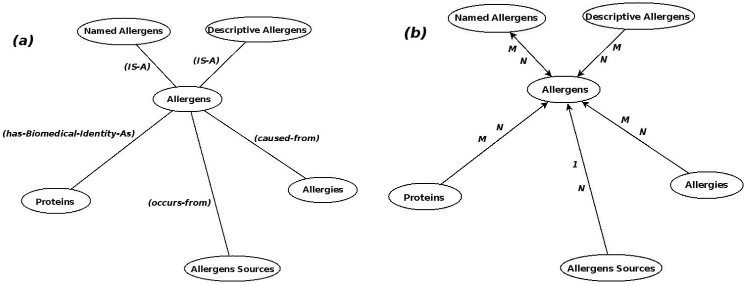
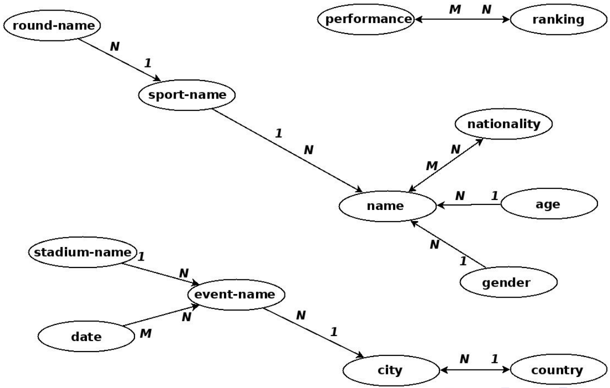


Figure: (a) The manually created ontology for the domain of allergens. (b) The automatically extracted ontology.

Experimental assessment for low-levels concepts

- We applied, on the corpus from the athletic domain, the proposed methodology, using a window of 50-symbols, for discovering semantic relations between low-level concepts.
- As low-level concepts we considered the thirteen different attributes used in the 20 high-level concepts. (56,494 instances' annotations)



Experimental assessment for low-levels concepts (cont'd)

- It is remarkable that the method also clusters the low-level concepts.
- The same results are also discovered for window size 100-symbols.
- For window size larger than 100-symbols, we observed that all the low-level concepts tend to be related with the more frequently occurring concept *name*.
- From experimentation we conclude that the best *WS* is related with the density of the annotated concept instances in the text.
  - The rule of thumb is: *the higher the density the lower the WS should be and vice versa*.

Conclusions

- We presented a novel method for discovering directed semantic relations for both high-level and low-level concepts.
- Our proposed method also finds cardinality restrictions for the instances of the discovered relations.
- We simply apply statistical methods to document metadata that is, to the location of concept instances in text.
- The proposed method was applied on two corpora of different domains and the results proved to be very promising in both domains

Future Plans

1. Use existing techniques for the automatic annotation of concepts' instances in order to further automate the proposed methodology
   - In the case of low-level concepts, named entity recognition techniques and also techniques which use the semantic-similarity among words will be employed.
   - In the case of high-level concepts, the work for the discovery of high-level concepts, performed in the context of the BOEMIE project will be examined.
2. We plan to extend our method, to support multiple inheritance.
3. Another aspect for future work is to apply the proposed approach in combination with already existing methods on relation discovery.

Thank you!

Questions...?

Results for semantic-correletion score of the low-level concepts

RELATION EXTRACTED $(round\_name \rightarrow sport\_name) = 0.610$

$P(round\_name - sport\_name) = 0.184$	$M(round\_name, sport\_name) = 2.303$	$Score = 0.610$
$P(round\_name - gender) = 0.188$	$M(round\_name, gender) = 2.195$	$Score = 0.602$
$P(round\_name - name) = 0.170$	$M(round\_name, name) = 2.0001$	$Score = 0.512$
$P(round\_name - nationality) = 0.120$	$M(round\_name, nationality) = 1.940$	$Score = 0.354$
$P(round\_name - ranking) = 0.101$	$M(round\_name, ranking) = 1.869$	$Score = 0.291$
$P(round\_name - date) = 0.084$	$M(round\_name, date) = 2.061$	$Score = 0.257$
$P(round\_name - performance) = 0.074$	$M(round\_name, performance) = 1.76$	$Score = 0.204$
$P(round\_name - event\_name) = 0.031$	$M(round\_name, event\_name) = 1.510$	$Score = 0.078$
$P(round\_name - age) = 0.016$	$M(round\_name, age) = 1.745$	$Score = 0.044$
$P(round\_name - city) = 0.017$	$M(round\_name, city) = 1.239$	$Score = 0.039$
$P(round\_name - country) = 0.006$	$M(round\_name, country) = 1.110$	$Score = 0.013$
$P(round\_name - stadium\_name) = 0.003$	$M(round\_name, stadium\_name) = 1.298$	$Score = 0.008$

Results for semantic-correletion score of the low-level concepts (cont'd)

RELATION EXTRACTED $(date \longrightarrow event\_name) = 0.587$

$P(date - event\_name) = 0.223$	$M(date, event\_name) = 1.632$	$Score = 0.587$
$P(date - city) = 0.167$	$M(date, city) = 1.489$	$Score = 0.416$
$P(date - name) = 0.103$	$M(date, name) = 1.054$	$Score = 0.212$
$P(date - country) = 0.072$	$M(date, country) = 1.445$	$Score = 0.177$
$P(date - sport\_name) = 0.083$	$M(date, sport\_name) = 1.110$	$Score = 0.175$
$P(date - ranking) = 0.081$	$M(date, ranking) = 1.044$	$Score = 0.166$
$P(date - nationality) = 0.079$	$M(date, nationality) = 1.031$	$Score = 0.161$
$P(date - performance) = 0.065$	$M(date, performance) = 0.981$	$Score = 0.129$
$P(date - gender) = 0.051$	$M(date, gender) = 1.020$	$Score = 0.104$
$P(date - stadium\_name) = 0.034$	$M(date, stadium\_name) = 1.533$	$Score = 0.087$
$P(date - age) = 0.021$	$M(date, age) = 1.132$	$Score = 0.045$
$P(date - round\_name) = 0.015$	$M(date, round\_name) = 1.332$	$Score = 0.036$