

# Recognizing Textual Entailment with Tree Edit Distance

Application to Question Answering

Milen Kouylekov

ITC-irst, Trento, Italy

30.11.2006

- 1 Language Variability and Textual Entailment
- 2 Distance-Based Approaches for Recognizing TE
- 3 Architecture for RTE
- 4 Experiments & Evaluation
- 5 Conclusions & Future Work

# Language Variability

- Definition: The same information can be expressed with different words and syntactic constructs.
- Language Variability problem has been attracting a lot of interest in Computational Linguistics during the years.
- The problem is usually approached in an application oriented manner.
- Language Variability is characterized by an extremely high degree of different phenomena.

# Language Variability

- Definition: The same information can be expressed with different words and syntactic constructs.
- Language Variability problem has been attracting a lot of interest in Computational Linguistics during the years.
- The problem is usually approached in an application oriented manner.
- Language Variability is characterized by an extremely high degree of different phenomena.
- Examples:
  - Lexical Variability: *Italy won the World Cup.*
  - Lexical Variability: *Squadra Azzura won the World Cup.*
  - Semantic Variability: *Italy became world champion for the fourth time.*
  - Syntactic & Semantic Variability: *The World Cup final was won by Italy.*

# Paraphrasing

- Popular approach to language variability.
- Definition: pairs of units with approximate conceptual equivalence (Barzilay 2003)
- Test: substituted for one another in many contexts.

# Paraphrasing

- Popular approach to language variability.
- Definition: pairs of units with approximate conceptual equivalence (Barzilay 2003)
- Test: substituted for one another in many contexts.
- Example:
  - Yahoo bought Overture.*
  - Yahoo purchased Overture.*
  - Yahoo pay for Overture.*
  - Yahoo completed acquisition of Overture.*

# Paraphrasing

- Popular approach to language variability.
- Definition: pairs of units with approximate conceptual equivalence (Barzilay 2003)
- Test: substituted for one another in many contexts.
- Example:
  - Yahoo bought Overture.*
  - Yahoo purchased Overture.*
  - Yahoo pay for Overture.*
  - Yahoo completed acquisition of Overture.*
- Does not provide a complete model of the problem of language variability:
  - Template: *X owned Y*
  - Sentence : **Datel corp.**  *sold today* **DT Communications**  *to* **Microsoft.**

# Textual Entailment

- General Framework proposed Dagan and Glickman (2004) addressing language variability.
- An Entailment Relation holds between two text fragments (i.e. text  $T$  and hypothesis  $H$ ) when the meaning of  $H$ , as interpreted in the context of  $T$ , can be inferred from the meaning of  $T$ .
- Directional - an expression entails the other, while the opposite may not.
- Probabilistic - the relation is not deterministic.

# Textual Entailment

- General Framework proposed Dagan and Glickman (2004) addressing language variability.
- An Entailment Relation holds between two text fragments (i.e. text  $T$  and hypothesis  $H$ ) when the meaning of  $H$ , as interpreted in the context of  $T$ , can be inferred from the meaning of  $T$ .
- Directional - an expression entails the other, while the opposite may not.
- Probabilistic - the relation is not deterministic.

- Example:

$T$  - *"For the first time in history, the players are investing their own money to ensure the future of the game," Atlanta Brave pitcher Tom Glavine said.*

$H$  - *Tom Glavine plays for the Atlanta Braves.*

# Entailment Rules

- A crucial role in textual entailment.
- Consists of an entailing template and an entailed template (right hand side RHS), which share the same variable scope.
- In order to apply an entailment rule, an appropriate prior or contextual (posterior) probability has to be assigned.

$$X \xleftarrow{\text{subj}} \text{sell} \xrightarrow{\text{obj}} Y \Rightarrow X \xleftarrow{\text{subj}} \text{own} \xrightarrow{\text{obj}} Y$$

# Recognizing Textual Entailment

- Textual Entailment was recently defined as a task by Dagan et.al. 2005.
- RTE takes as input a  $T-H$  pair and consists in automatically determining whether an entailment relation between  $T$  and  $H$  holds or not.
- Evaluated in two evaluation campaigns, the Pascal Recognizing Textual Challenge 1 & 2.
  - Attracted a lot attention in the scientific community.
  - The organizers provided a set of  $T-H$  pairs split into positive and negative examples.
  - Pairs were collected by human annotators from seven different application scenarios.
  - The system performance ranged from 50% to 60% for RTEC 1 and from 53% to 75% for RTEC 2.

# Recognizing Textual Entailment

## Requirements:

- Modular in order to combine different levels of linguistic processing.
- Flexible enough to use and test different sources of knowledge.
- Intrinsically indeterministic.
- Direct evaluation (on the task itself) and indirect evaluation (in application scenarios).

## Solutions:

- The edit distance framework.
- Edit operations express different sources of knowledge for resolving language variability problems
- We assign an entailment relation if the overall cost of the transformation is below a certain threshold, empirically estimated on the training data.

# Distance Based Approach for RTE

An edit distance approach for RTE assumes that the distance between  $T$  and  $H$  is a characteristic that separates the positive pairs from the negative pairs.

- 1 It exists a function, with range from 0 to  $K$ , that calculates an entailment score of a  $T$ - $H$  pair based on the edit distance between  $T$  and  $H$ .
- 2 If  $T$  and  $H$  are the same, then  $T$  entails  $H$ .
- 3 If  $T$  and  $H$  are completely different then,  $T$  does not entail  $H$ .
- 4 It exists a distance boundary (threshold)  $S$ ,  $0 < S < K$ , that separates the positive from the negative examples.

# Edit Operations

We assume that the distance between  $T$  and  $H$  is computed as the cost of the editing operations on text fragments that transform  $T$  into  $H$ .

**Edit Operation** - *An operation that converts a text fragment  $A$  into another text fragment  $B$  ( $A \rightarrow B$ ) with a certain cost  $\gamma(A \rightarrow B)$ .*

- 1 **Insertion**  $\Lambda \rightarrow A$ : Inserts a text fragment  $A$  from  $H$  in  $T$ .
- 2 **Deletion**  $A \rightarrow \Lambda$ : Removes a text fragment  $A$  from  $T$ .
- 3 **Substitution**  $A \rightarrow B$ : Replaces a text fragment  $A$  from  $T$  with a text fragment  $B$  from  $H$ .

We define an entailment score function in the following way:

$$\text{score}_{\text{entailment}}(T, H) = \frac{\gamma(T, H)}{\gamma_{\text{nomap}}(T, H)}$$

# Edit Operations Cost

The cost of the edit operations is defined in the following way:

$$\begin{aligned}\gamma(\Lambda \rightarrow A) &= \text{const}_i; \\ \gamma(A \rightarrow \Lambda) &= \text{const}_d \\ \gamma_{i+d}(A \rightarrow B) &= \gamma(\Lambda \rightarrow A) + \gamma(A \rightarrow \Lambda) \\ \gamma(A \rightarrow B) &= \begin{cases} 0 & A = B \\ \gamma_{i+d}(A \rightarrow B) & \text{otherwise} \end{cases}\end{aligned}$$

# Lexical Entailment Rules

Our approach to use lexical entailment rules to approximate the cost of substitution.

$$\gamma(A \rightarrow B) = \begin{cases} 0 & A = B \\ \gamma_{i+d}(A \rightarrow B) * (1 - P_{correct}(A \Rightarrow B)) & A \Rightarrow B \\ \gamma_{i+d}(A \rightarrow B) & \textit{otherwise} \end{cases}$$

# Levenshtein Distance

Named after Vladimir Levenshtein, who introduced this distance in 1965.

- Measure of the similarity between two strings.
- Minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character.
- Used in spell checkers, speech recognition, DNA analysis and plagiarism detection.

# Levenshtein Distance

Named after Vladimir Levenshtein, who introduced this distance in 1965.

- Measure of the similarity between two strings.
- Minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character.
- Used in spell checkers, speech recognition, DNA analysis and plagiarism detection.

For RTE task we convert the text of the sentences  $T$  and  $H$  into sequences of words.

- **Insertion**  $\Lambda \rightarrow A$ : insert a word  $A$  from  $H$  into  $T$ .
- **Deletion**  $A \rightarrow \Lambda$ : delete a word  $A$  from  $T$ .
- **Substitution**  $A \rightarrow B$ : substitute a word  $A$  from  $T$  with a word  $B$  from  $H$ .

# Example

*T - A Union Pacific freight train hit five people.*

*H - A Union Pacific freight train struck five people.*

$$\gamma(\Lambda \rightarrow A) = 1$$

$$\gamma(A \rightarrow \Lambda) = 1$$

$$\gamma(A \rightarrow B) = \begin{cases} 0 & A = B \\ 2 & \textit{otherwise} \end{cases}$$

Edit Distance is equal to 2, NoMap distance is equal to 16,  
Entailment Score is equal to 0.125.

# Examples

*T - Jennifer Hawkins is the 21-year-old beauty queen from Australia.*

*H - Jennifer Hawkins is Australia's 21-year-old beauty queen.*

*T - CD Technologies announced that it has closed the acquisition of Datel, Inc.*

*H- Datel Inc was acquired by C&D Technologies.*

# Tree Edit Distance

Tree edit distance algorithm on the syntactic representations of both  $T$  and  $H$ .

- The tree edit distance algorithm described by Zhang and Shasha 1990.
- Edit operations (Insertion, Deletion, Substitution) are allowed on single nodes only.
- Node order is relevant: nodes are re-arranged according to: subj, obj, mods.
- The original algorithm does not consider labels on edges: relations names are concatenated to node names.

$$eat \xrightarrow{\text{subj}} John \Rightarrow eat \rightarrow John\#subj$$



# Extension: Constraints

We extended the substitution cost function with a mechanism of control.

- A pattern applied to the dependency tree of either  $T$  and  $H$ .
- Two types of constraints: positive and negative

# Extension: Constraints

We extended the substitution cost function with a mechanism of control.

- A pattern applied to the dependency tree of either  $T$  and  $H$ .
- Two types of constraints: positive and negative

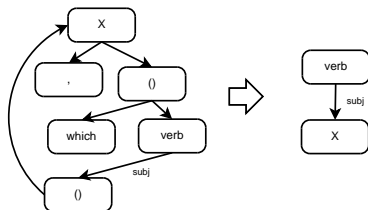
$$\gamma(A \rightarrow B) = \begin{cases} 0 & A = B \\ 0 & A \neq B \ \& \ \text{cons}_{\text{pos}}(A, B) \\ \gamma_{i+d}(A \rightarrow B) & A = B \ \& \ \text{cons}_{\text{neg}}(A, B) \\ \gamma_{i+d}(A \rightarrow B) & \text{otherwise} \end{cases}$$

# Positive Constraints

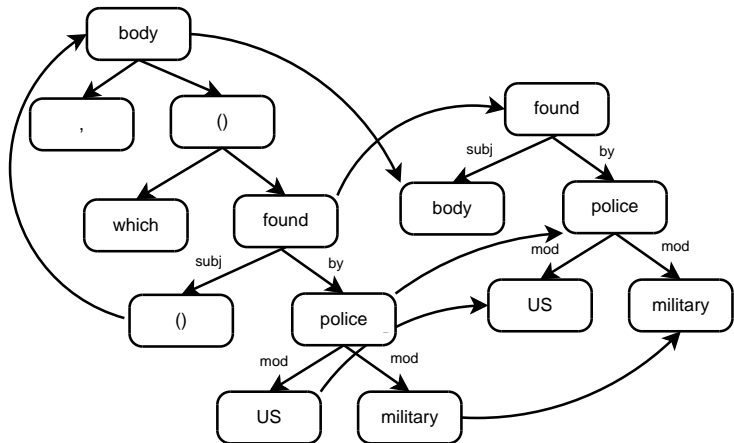
- Model syntactic language variability: relative clause, apposition, conjunction, transparent head.
- Encoded as small set of parser dependent transformation rules.
  - Pair of syntactic templates with variables that are strictly matched to a subtree of  $T$  and the corresponding subtree of  $H$ .
  - Activated if both templates can be completely matched.

# Positive Constraints

- Model syntactic language variability: relative clause, apposition, conjunction, transparent head.
- Encoded as small set of parser dependent transformation rules.
  - Pair of syntactic templates with variables that are strictly matched to a subtree of  $T$  and the corresponding subtree of  $H$ .
  - Activated if both templates can be completely matched.



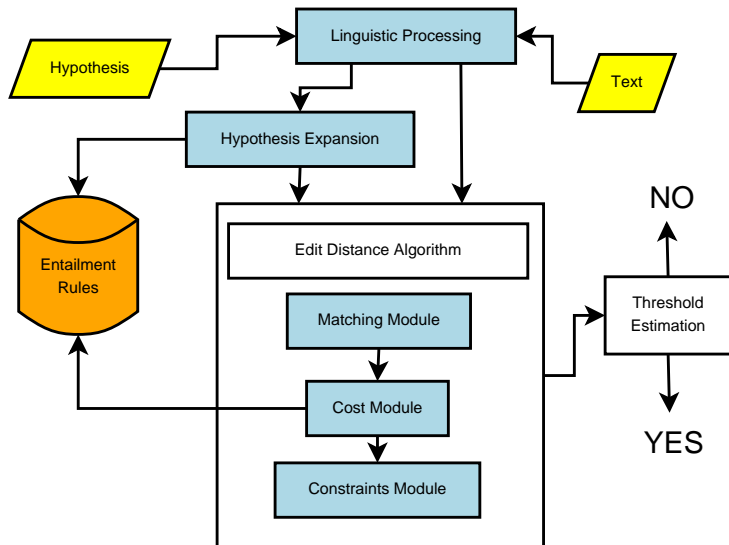
# Example



# Negative Constraints

- Deleted words from  $T$  usually influence the meaning of already matched (substituted) words.
- Marneffe et. al. 2006 - different features that can cause mismatching between  $T$  and  $H$ .
  - Polarity features - negation (*not*), downward-monotone quantifiers (*no*, *few*), restricting prepositions (*without*, *except*) and superlatives (*tallest*).
  - Antonymy features - (*hot soup vs cold soup*).
  - Modality features - simple patterns for modal reasoning (i.e. *must*, *maybe*).
  - Quantifier features. *every company vs company*.

# RTE System Architecture



- Substitution - Lexical Entailment Rules
  - WordNet - lexical chains.
  - Similarity thesaurus - word similarities.
- Hypothesis Expansion
  - TEASE - Web-based acquisition of entailment rules.
  - Dirt - Corpus-based acquisition of inference rules.
- Insertion
  - Text Corpus for calculating relevance of words (inverse document frequency).

# Lexical Entailment Rules

## WordNet

- WordNet is a lexical database which includes lexical and semantic relations among word senses.
- WordNet as paraphrasing resource is discussed by Moldovan and Rus 2001.
- We have defined entailment rules over WordNet considering a subset of the relations among synsets: synonym, hypernym, entails and pertains.

## Word Similarity Thesaurus (UniAlberta Thesaurus)

- Developed in University of Alberta by Dekang Lin 1998.
- A thesaurus that calculates similarity between words based on dependency relations.
- For each word, the thesaurus lists up to 200 most similar words and their similarities, estimated on a parsed corpus using frequency counts of the dependency triples.

# Syntactic Entailment Rules

## DIRT System Lin and Pantel 2001

- A corpus based approach for discovering inference rules.
- Improve the performance of a Question Answering system.
- They define an inference rule as a couple of *similar paths* between nodes in two dependency trees.
- Extended version of the Distributional Hypothesis (Harris 1985)  
- two texts tend to occur in a similar contexts, then the meanings of the syntactic paths tend to be similar.

## TEASE System Szpektor et.al. 2004

- Web based and unsupervised acquisition of entailment relations.
- The aims: 1) cover the broadest possible range of meanings; 2) keep the extracted templates as general as possible.

# Insertion Cost Functions

The intuition underlying insertion is that its cost is proportional to the relevance of the word  $w$  to be inserted (i.e., inserting an informative word has a higher cost than inserting a less informative word).

- Inverse document frequency (idf) - the most popular measure in Information Retrieval Community.

$$rel(w) = idf(w) = \log \frac{N}{N_w}$$

- The words with higher position in the tree (i.e. closer to the root of the tree) are considered more relevant to the meaning expressed by a certain phrase.

$$rel(w) = 10 - \#parents\_of\_w$$

# RTE Results

	Pascal 1	Pascal 2
Baseline	0.550	0.560
IDF	0.559	0.561
parents	0.573	0.577
UniAlbera thesaurus	0.560	0.564
WordNet	0.572	0.580
Dirt	0.578	0.585
Combined	0.584	0.613

# Evaluation on Question Answering

- Answer validation is a crucial step in the Question Answering loop.
- Different techniques: lexical syntactic overlapping, logical form representation and reasoning, statistical web validation.
- Answer Validation as an RTE task:

*A candidate answer  $ca$  to a question  $q$  is a correct answer if and only if the text from which  $ca$  is extracted entails the affirmative form of  $q$ , with  $ca$  inserted into the appropriate position.*

# Answer Validation Exercise

- New evaluation exercise - Penas et. al. 2006 part of the Cross Language Evaluation Forum (CLEF).
- The goal is to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by QA systems.
- Participant systems receive a set of pairs  $T$ - $H$  built from the QA main track responses of the participating systems in CLEF 2006.
- $H$  is constructed semi-automatically from the question in affirmative form with the answer replaced in the proper position.

# AVE Participation Results

Precision(YES)	Recall(YES)	F measure
0.3025	0.5023	0.3776

- More difficult than the Pascal RTE.
- Real Examples: Need additional processing.
- Type Distinction (Factoid vs Definition)
- Answer Validation vs Answer Filtering.

# Conclusions

- The major result of our work is a distance-based framework for recognizing textual entailment.
- We have investigated and modeled different lexical and syntactic variability phenomena.
- Lexical and syntactic entailment rules can significantly improve the performance of a distance-based system for RTE
- Our conclusion from this evaluation is that RTE is a good approach for the QA.
- More work is required to solve problems: discourse phenomena (anaphora resolution) and basic syntactic variability, like transparent heads.

Language variability is expressed with complex linguistic phenomena.

- Space and temporal reasoning, in particular processing of temporal anaphora and numerical expressions.
- Discourse processing, such as anaphora resolution and text structure theory to connect text fragments.
- World knowledge, where the use of facts about the state of the world can allow to make complex reasoning.