Local Alignment Kernels for Relation Extraction

Sophia Katrenko Pieter Adriaans

University of Amsterdam, the Netherlands

Seventh International Tbilisi Symposium on Language, Logic and Computation, 2007

Outline

Examples

Generic Relations

- Mary looked back and whispered: "I know every tree in this forest, every scent". (Part-Whole)
- A person infected with a particular flu virus strain develops antibody against that virus. (Cause-Effect)
- The apples are in the basket. (Content-Container)

Examples

Generic Relations

- Mary looked back and whispered: "I know every tree in this forest, every scent". (Part-Whole)
- A person infected with a particular flu virus strain develops antibody against that virus. (Cause-Effect)
- The apples are in the basket. (Content-Container)

Domain-Specific Relations

- The expression of rsfA is under control of both sigma(F) and sigma(G).
- Therefore, the role of sigmaB-dependent katX expression remains obscure.

What are relations useful for?

1. Information extraction systems
2. Hypothesis generation (D. Swanson, 1986)
3. Question answering (Ch. Lee et al., 2007; R. Srihari, 1999)
4. . . .

Representations and Methods

1. Different Representations
   - Subsequences in the sentences (Bunescu et al., 2005)
   - Syntactic structures
     - dependency paths (Bunescu et al., 2005)
     - pre-defined levels (Katrenko et al., 2006)

Representations and Methods

1. Different Representations
   - Subsequences in the sentences (Bunescu et al., 2005)
   - Syntactic structures
     - dependency paths (Bunescu et al., 2005)
     - pre-defined levels (Katrenko et al., 2006)
2. Different Methods
   - hand-written patterns (Hearst, 1992)
   - kernel methods (Zelenko et al., 2003; Zhao et al., 2005; Culotta and Sorensen, 2004)
   - pattern induction methods (Snow at al., 2005)
   - other ML methods depending on the data representation

Our choice

1. Step I: fix a representation
   - Dependency paths, i.e. any relation mention $e = (x, y)$ is presented as $e = (x \rightarrow z_1 \rightarrow \ldots \rightarrow z_n \rightarrow y)$ and our goal is to find a hypothesis H, $H : E \rightarrow \{0, 1\}$ where E is a set of positive and negative examples of a given relation

Our choice

1. Step I: fix a representation
   - Dependency paths, i.e. any relation mention $e = (x, y)$ is presented as $e = (x \rightarrow z_1 \rightarrow \ldots \rightarrow z_n \rightarrow y)$ and our goal is to find a hypothesis H, $H : E \rightarrow \{0, 1\}$ where E is a set of positive and negative examples of a given relation
2. Step II: fix a method
   - kernel methods (but used a bit differently)

A *very* short intro to kernel methods

1. Kernel methods (KM) are an alternative (Vapnik, 1998) to the feature-based representation
2. KM retain the original representation of the objects and compute a similarity function between a pair of objects

Definition 1

Let X be a set and $K : X \times X \rightarrow \Re$. K is a kernel on $X \times X$ if K is symmetric and positive definite (for any $N \geq 1$ and any $x_1, \ldots x_N \in X$, the matrix X defined by $K_{ij} = K(x_i, x_j)$ is positive definite, i.e. $\sum_{ij} c_i c_j K_{ij} \geq 0$ for all $c_1, \ldots, c_N \in \Re$)

A *very* short intro to kernel methods

1. Kernel methods (KM) are an alternative (Vapnik, 1998) to the feature-based representation
2. KM retain the original representation of the objects and compute a similarity function between a pair of objects

Definition 1

Let X be a set and $K : X \times X \rightarrow \Re$. K is a kernel on $X \times X$ if K is symmetric and positive definite (for any $N \geq 1$ and any $x_1, \ldots x_N \in X$, the matrix X defined by $K_{ij} = K(x_i, x_j)$ is positive definite, i.e. $\sum_{ij} c_i c_j K_{ij} \geq 0$ for all $c_1, \ldots, c_N \in \Re$)

3. Kernel computes an inner product

$$< \boldsymbol{x}, \boldsymbol{y} >= \sum_{i=1}^{n} x_i y_i$$

by implicitly mapping the examples to the feature space

Similarity measures

1. Kernel methods proved to be accurate but can we do better? Why not use more elaborate measures?
   - Biologists
     - Smith-Waterman distance on two sequences of amino acids (Smith and Waterman, 1981)
   - Linguists (based on Cohen et al., 2003)
     - term-based (TF-IDF)
     - edit distance (Levenshtein distance, Smith-Waterman)
     - HMM-based

Smith-Waterman distance

- Given two sequences $\boldsymbol{x} = x_1 x_2 \ldots x_n$ and $\boldsymbol{y} = y_1 y_2 \ldots y_m$, Smith-Waterman distance is defined as the local alignment score of their best alignment, or in dynamic programming setting

Definition 2 (Smith-Waterman distance)

$$SW(i,j) = max \begin{cases} 0 \\ SW(i-1, j-1) + d(x_i, y_j) \\ SW(i-1, j) - G \\ SW(i, j-1) - G \end{cases}$$

- G is a penalty gap and $d(x_i, y_j)$ is a substitution score
- SW score can be computed in $O(n * m)$ time

Example

Example

$G = 1$, $d(x, x) = -2$, $d(x, y) = 1(x \neq y)$

		T	B	I	L	I	S	I
		0	0	0	0	0	0	0
T	0	2	1	0	-1	-1	-1	-1
I	0	1	1	3	2	1	0	1
F	0	0	0	2	2	1	0	0
L	0	-1	-1	1	4	3	2	1
I	0	-1	-2	1	3	6	5	4
S	0	-1	-2	0	2	5	**8**	7

Example

Example

$G = 1$, $d(x, x) = -2$, $d(x, y) = 1(x \neq y)$

		T	B	I	L	I	S	I
		0	0	0	0	0	0	0
T	0	2	1	0	-1	-1	-1	-1
I	0	1	1	3	2	1	0	1
F	0	0	0	2	2	1	0	0
L	0	-1	-1	1	4	3	2	1
I	0	-1	-2	1	3	6	5	4
S	0	-1	-2	0	2	5	**8**	7

T B I - L I S I
T - I F L I S -

Local alignment kernel (1)

1. How can we define a kernel function based on the local alignment score?
   - A kernel must be valid
   - Using an original SW score does not result in a valid kernel because it keeps the contribution of the best local alignment to quantify the similarity between two sequences (does not sum up the contribution of all possible local alignments)

Local alignment kernel (1)

1. How can we define a kernel function based on the local alignment score?
   - A kernel must be valid
   - Using an original SW score does not result in a valid kernel because it keeps the contribution of the best local alignment to quantify the similarity between two sequences (does not sum up the contribution of all possible local alignments)

   Solution
   - Kernel becomes valid if it is defined as follows (Vert et al., 2004)

$$K_{LA}(x, y) = \sum_{\pi \in A(x,y)} exp^{\beta s(x,y,\pi)}$$

   where $s(x, y, \pi)$ is a score of the local alignment π from the set of all possible alignments A.

Local alignment kernel (2)

SW and LAK are related in the following way:

$$\lim_{\beta \to \infty} \frac{1}{\beta} K_{LA}(x, y) = SW(x, y)$$

How to calculate $d(\bullet, \bullet)$ in SW?

Biologists

- predefined blossum matrix

Several options widely used in NLP

- statistical measures (semantic similarity given a large corpus)
- measures defined over various semantic resources such as WordNet

Distributional hypothesis

Distributional similarity (Firth, 1957; Harris, 1968)

Words found in the similar contexts tend to be semantically similar

Mohammed and Hirst, 2005

Distributionally similar words tend to be semantically similar, where two words w_1 and w_2 are said to be distributionally similar if they have many common co-occurring words and these co-occurring words are ech related to w_1 and w_2 by the same syntactic relation.

Dice measure

Dice measure is defined as follows

$$dice(w_1, w_2) = \frac{W_1 \cap W_2}{W_1 \cup W_2}$$

where W_1 and W_2 are sets whose members co-occur with w_1 and w_2 respectively.

Settings...

1. Data
   - Training set from "Learning Language in Logic" workshop containing interactions between proteins and genes (subset of Medline)
   - biomedical journals for estimating distributional similarity (from TREC 2007)
2. Dice measure
   - 10,000 occurrences w_1 and w_2
   - context of two tokens to the left and to the right
3. LAK parameters
   - Gap penalty $G = 2$

Results (1)

1. Given string kernel (SK) as a baseline, LAK outperforms SK by approx. 30% (accuracy of LAK 92,10%, string kernels - 63,59%.
2. It also performs better than methods working on level-based representation (91,32%).

Results (2)

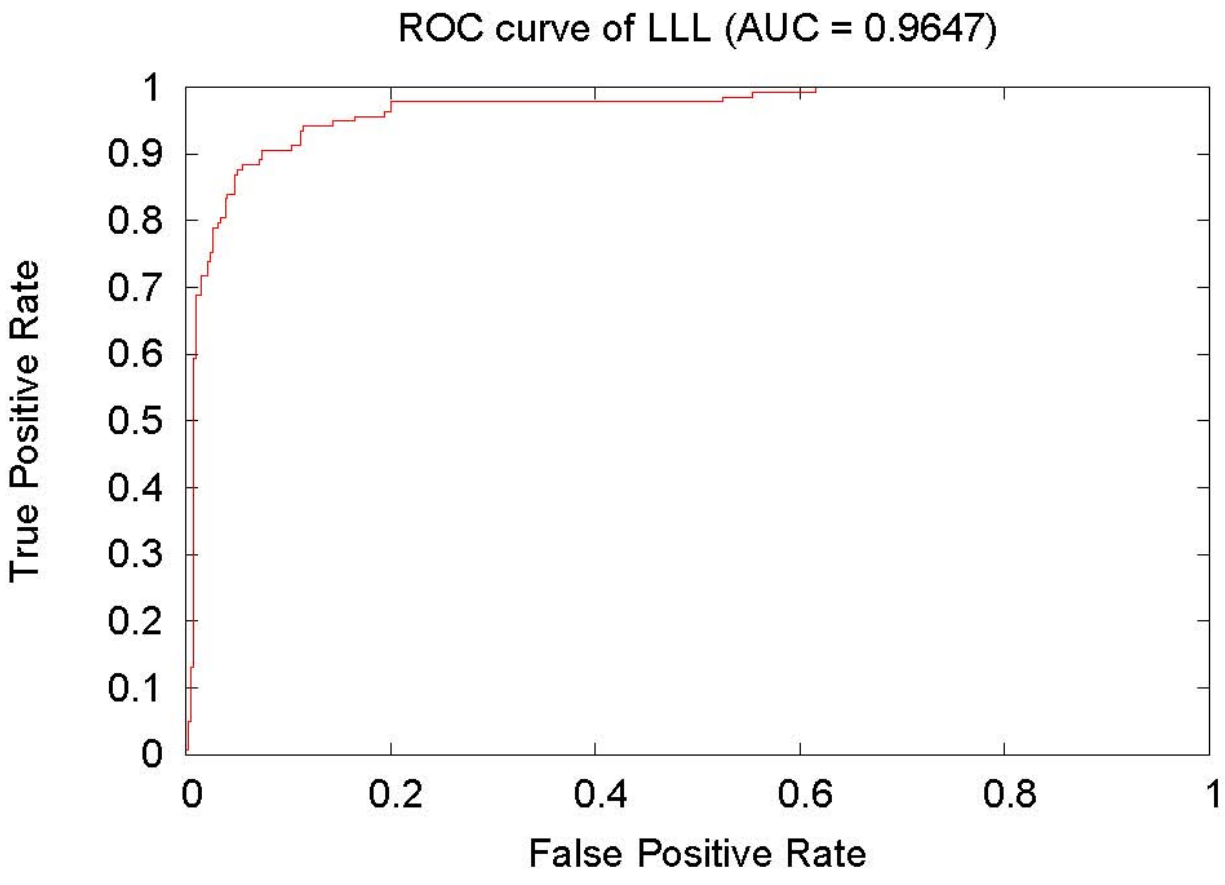

Figure: LAK on LLL (10-fold cross-validation)

Future Work

1. A hypothesis of LAK handling well the data sparseness
2. Other statistical measures (Jaccard, cosine, etc.)
3. Measures calculated on syntactic functions rather than immediate context
4. Experiments in other domains (or more generic, e.g. *part-whole* relation)

Summary

- . . . presented a novel method based on the local alignment of sequences
- . . . put together measures of distributional relatedness and similarity measures defined on sequences
- . . . presented some promising results on the relation extraction task in the biomedical domain