Using Global Constraints and Reranking to Improve Cognates Detection

Michael Bloodgood[1] Benjamin Strauss[2]

[1]Department of Computer Science
The College of New Jersey

[2]Computer Science and Engineering Dept.
The Ohio State University

Introduction

Background Cognates detection is the task of identifying words across languages that have a common origin. Cognates are used for protolanguage reconstruction and cross-language dictionary lookup. Cognates can also improve the quality of machine translation, word alignment, and bilingual lexicon induction.

Current Solution Create a score matrix for all word pairs based on weighted combinations of component scores. The component scores are computed on the basis of word context information, word frequency information, temporal information, word burstiness information, and phonetic information.

Our Contribution We propose a new algorithm for rescoring the matrix by taking into account scores assigned to other word pairs. Precision and recall are improved by large amounts in experiments across three language pairs with both traditional and new large data testing conditions.

Motivation

Spanish \ Portuguese	⋯	cozinhar	⋯	andar
⋮	⋱			
caminar		.80 (2) .40 (2) .20		.70 (1) .70 (1) .70
⋮			⋱	
cocinar		.99 (1) .99 (1) .99		.20 (2) .10 (2) .05

Blue - Score from Initial Score Matrix

(#) - Reverse Rank

Green - Score after Reverse Rank

(#) - Forward Rank

Black - Score after Forward Rank

General Task

Let $X = \{x_1, x_2, ..., x_n\}$. Let $Y = \{y_1, y_2, ..., y_n\}$.
Extract (x, y) pairs such that (x, y) are in some relation R.

General Algorithm

Score Matrix

$$Score_{X,Y} = \begin{bmatrix} s_{x_1,y_1} & \cdots & s_{x_1,y_n} \\ \vdots & \ddots & \vdots \\ s_{x_n,y_1} & \cdots & s_{x_n,y_n} \end{bmatrix} \quad (1)$$

Rescoring

Reverse Rank

$$reverse\_rank(x_i, y_j) = |\{x_k \in X | s_{x_k,y_j} \geq s_{x_i,y_j}\}| \quad (2)$$

Score RR

$$score_{RR}(x_i, y_j) = \frac{s_{x_i,y_j}}{reverse\_rank(x_i, y_j)} \quad (3)$$

Forward Rank

$$forward\_rank(x_i, y_j) = |\{y_k \in Y | s_{x_i,y_k} \geq s_{x_i,y_j}\}| \quad (4)$$

Combining Reverse Rank and Forward Rank

1-Step Approach

$$score_{RR\_FR\_1step}(x_i, y_j) = \frac{s_{x_i,y_j}}{product}, \quad (5)$$

where

$$product = reverse\_rank(x_i, y_j) \times forward\_rank(x_i, y_j). \quad (6)$$

2-Step Approach

$score_{RR\_FR\_2step}$ involves first computing the reverse rank and re-adjusting every score based on the reverse ranks. Then in a second step the new scores are used to compute forward ranks and then those scores are adjusted based on the forward ranks.

Maximum Assignment

We used the Hungarian Algorithm to optimize:

$$\max_{Z \subseteq X \times Y} \sum_{(x,y) \in Z} score(x, y)$$
$$\text{s.t. } (x_i, y_j) \in Z \Rightarrow (x_k, y_j) \notin Z, \forall k \neq i$$
$$(x_i, y_j) \in Z \Rightarrow (x_i, y_k) \notin Z, \forall k \neq j. \quad (7)$$

Computation of Initial Score Matrix

Lemmatization

- English - NLTK WordNetLemmatizer [Bird et al., 2009]
- French, German, Spanish - TreeTagger [Schmid, 1994]

Word Context Information 2012 Google 5-gram corpus for English, French, German, and Spanish [Michel et al., 2010]

Frequency Information Computed using the same corpora as Word Context Information.

Temporal Information Computed using the following corpora:

- English - English Gigaword Fifth Edition
- French - French Gigaword Fifth Edition
- Spanish - Spanish Gigaword Fifth Edition
- German - Web crawling and extracting news articles from http://www.tagesspiegel.de/

Word Burstiness Information Computed using the same corpora as Temporal Information.

Phonetic Information Computed using a measurement based on Normalized Edit Distance (NED).

Combining Information Sources For each candidate cognate pair (x, y), its final score is:

$$score(x, y) = \sum_{m \in metrics} w_m score_m(x, y), \quad (8)$$

where $metrics$ is the set of measurements listed above; w_m is the learned weight for metric m; and $score_m(x, y)$ is the score assigned to the pair (x, y) by metric m.

Experiments

We used the following language pairs:

- French-English
- German-English
- Spanish-English

We used the following testing conditions:

- Small Data (Traditional)
- Large Data (New)

We used the following as our baseline:

- Initial Score Matrix without any rescoring

Using Global Constraints to Rescore - Results

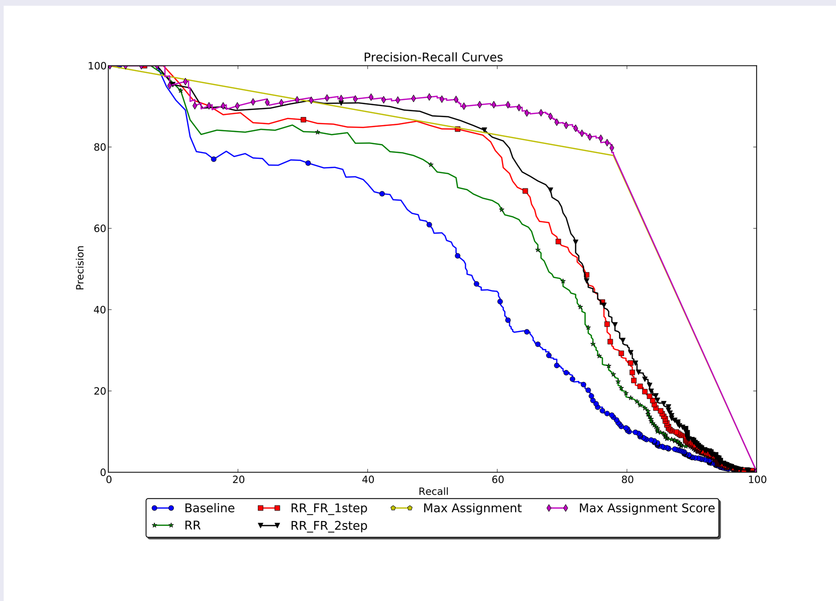

Figure: Precision-Recall Curves for French-English (small data)

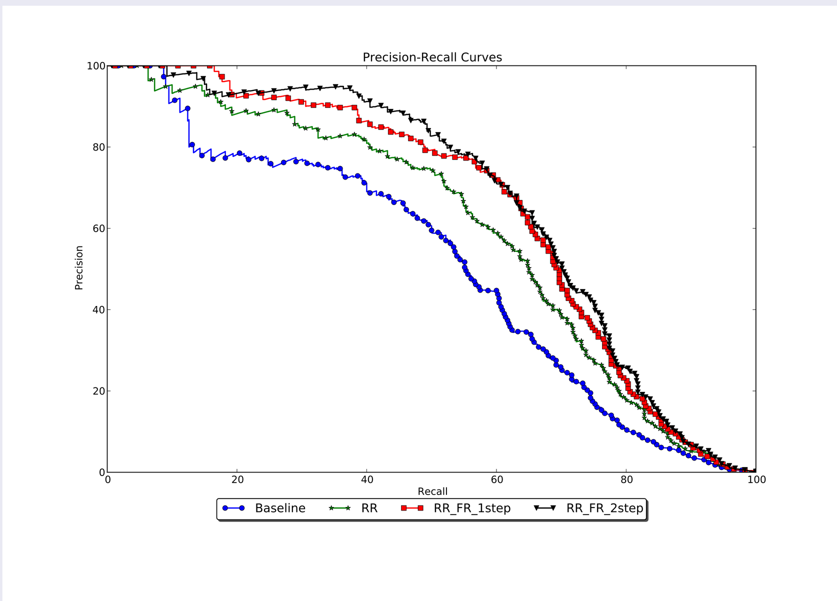

Figure: Precision-Recall Curves for French-English (large data)

Method	Max F1	11-point IAP
Baseline	54.92	50.99
RR	62.94	59.62
RR_FR_1step	68.35	64.42
RR_FR_2step	69.72	67.29

Table: French-English Performance (small data)

Method	Max F1	11-point IAP
Baseline	55.08	51.35
RR	60.88	58.79
RR_FR_1step	65.87	63.55
RR_FR_2step	65.76	65.26

Table: French-English Performance (large data)

Conclusion

We presented new methods for rescoring a matrix of initial scores and new large data testing conditions for evaluation. Our new methods are complementary to existing state of the art methods, easy to implement, computationally efficient, and effective in improving performance.

Acknowledgment

We would like to thank TCNJ students Ethan Kochis and Garrett Beatty for their help with building this poster.

Bibliography

Bird, S., Klein, E., and Loper, E. (2009).
Natural Language Processing with Python.
O'Reilly Media, Inc., 1st edition.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2010).
Quantitative analysis of culture using millions of digitized books.
Science, 331(6014):176–182.

Schmid, H. (1994).
Part-of-speech tagging with neural networks.
In *COLING*, pages 172–176.
http://www.aclweb.org/anthology/C/C94/C94-1027.pdf.