Alignment-Based Discriminative String Similarity

Shane Bergsma and Greg Kondrak
University of Alberta
ACL 2007
String Similarity

- Input: Pair of Strings
- Output: Measure of Similarity
- Our approach: discriminative, data-driven
- Features are substrings extracted from a character-based alignment of the strings
- Evaluate on cognate identification
- Excellent results
Outline

1. String similarity and its applications
2. Previous approaches
3. Alignment-based discriminative similarity
4. Cognate Data Generation
5. Experiments and Results
1. String Similarity

- **Example: Spelling Correction:**

|           |               |
|-----------|---------------|
| Wallmart  | Wall-mart     |
| Britany Spears | Britney Spears |
| Amtrack   | Amtrak        |
| Hillary Duff | Hilary Duff   |
| Geneology | Genealogy     |

(From Yahoo’s most common search query spelling errors)
String Similarity

- Example: Word Alignment

```
startup properties and options
```

- Words with similar form and meaning are called cognates:

```
propriétés de démarrage et options
```

```
propriétés properties
options options
```
Cognate Identification

• Cognates:
  – Ancestral: English/German night/nacht
  – Borrowed: English/Japanese trampoline/toranporin

• Our focus: “translational” cognates

• String similarity indicates how likely the two words are to be translations based on their orthographic similarity
2. Previous Approaches

Traditional Approaches:

• Normalized Edit Distance
• Longest Common Subsequence Ratio (LCSR)

– Efficient, no training data needed, but not optimized for specific tasks
Improved Approaches

• Tiedemann (1999), Mulloni & Pekar (2006)
  – Look for consistent spelling changes across cognates: e.g. English/German *electric*- *elektrisch*
  – Re-weight LCSR/NED to add uncounted “mutations”
Klementiev and Roth (2006)

• Originally for Named-Entity Transliteration
• Discriminative String Similarity:
  – Extract features for pairs of strings: create feature vector
  – Label feature vectors as positive or negative
  – Train classifier on labelled feature vectors
Klementiev and Roth (2006)

For Cognates:

• E.g: Japanese/English *sutoresu:*stress

• *sutoresu* → \{ s, u, t, o, r, e, s, u, su, ut, ... \}

• *stress* → \{ s, t, r, e, s, s, st, tr, re, es, ss \}

• Gives features:

\{s-s, s-t, s-st, su-s, su-t, su-st, su-tr... r-s, r-s, r-es ... \}
3. Alignment-Based Discriminative

• The character-based alignment generates the features for discriminative learning:

  $\{\wedge s u t o r e s u, \wedge s t r e s s, s-s, s-u-s, u-t-t, t-t, \ldots, e-s-e-s, s-s, s-u-s-s \ldots\}$

• Gives features:

• Creates a more focused feature space for a given max substring size
Alignment-Based Discriminative

- Include other features like NED and special longer phrases
- Learn classifier with SVM, score by positive distance from SVM hyperplane.
  - See paper for more details
Outline

1. String similarity and its applications
2. Previous approaches
3. Alignment-based discriminative similarity
4. Cognate Data Generation
5. Experiments and Results
4. Cognate Data Generation

• Manual:
  – Get linguist (or computational linguist) to identify all cognates

• Automatic:
  – Define cognates to be all pairs with \( \text{LCSR} \geq 0.58 \) that have the same meaning (Melamed, 1999).
Cognate Data Generation

• Determining common meaning:
  – Method 1:
    • Are they translations in a translation lexicon?
  – Method 2:
    • Are they commonly aligned in a word-aligned bitext?
Cognate Data Generation

• For a given foreign word $f$, find cognates among $E_f$ that have LCSR $\geq 0.58$

  – Examples:

| Language | Foreign word $f$ | Cognates $E_{f+}$ | False Friends $E_{f-}$ |
|----------|------------------|-------------------|------------------------|
| Japanese | napukin          | napkin            | nanking, pumpkin, snacking, sneaking |
| French   | abondamment      | abundantly        | abandonment, abatement, wonderment |
Cognate Data Generation

- Not a ranking task – not every foreign word has a cognate
- Rather, a pairwise classification:
  + napukin, napkin
  - napukin, nanking
  - napukin, pumpkin
- Note: automatically creates competitive counter-examples for learning
5. Experiments and Results

1) Bitext Experiments
   – French-English, Spanish-English, German-English
   – Word-aligned data from the Europarl corpus

2) Dictionary Experiments
   – Word pairs from www.Freelang.net
   – French-English, Spanish-English, German-English, Greek-English, Japanese-English, Russian-English
   – Romanization of Greek, Russian
String Similarity Performance

11-pt Average Precision (%)

- Fr
- Es
- De
- Fr
- Es
- De
- Gr
- Jp
- Rs

Bitext ➔ Dictionary
String Similarity Performance

- LCSR
- Tiedemann
- Klementiev & Roth

11-pt Average Precision (%)

Fr - Es - De - Fr - Es - De - Gr - Jp - Rs

Bitext  Dictionary
String Similarity Performance

- LCSR
- Tiedemann
- Klementiev & Roth
- Alignment-Based Discriminative

11-pt Average Precision (%)
Bitext Fr-En Learning Curve
## Important Features

| Language     | Feature  | Weight | Example               |
|--------------|----------|--------|-----------------------|
| French       | ées-ed   | +8.0   | vérifiéées:verified    |
| German       | k-c      | +5.5   | kreativ:creative      |
| Greek        | f-ph     | +4.1   | symfonia:symphony     |
| Japanese     | ou-ou    | -2.6   | handoutai:handout*    |
| Spanish      | mos-s    | -5.1   | toleramoss:tolerates* |
Conclusion

• First approach to apply discriminative string similarity to cognate identification
• Alignment-based features allow for strong gains in performance
• Phonetic, syntactic or semantic features can be incorporated into this framework
Thanks
String Similarity

- Example: Named Entity Transliteration:

| English NE | Russian NE |
|------------|-----------|
| lilic      | лилич    |
| fletcher   | флетчер   |
| bradford   | брэдфорд |
| isabel     | изабель   |
| hoffmann   | гофман    |
| kathmandu  | катманду |

(From Klementiev & Roth (2006))
Brill and Moore (2000)

- Get probability of edit operations for spelling correction
- Expand non-match substitutions with adjacent edits
- Learn generative model with EM

```
actual
/    \
|     |
|     |
akgsual
```

```
a → a, c → k, e → g, t → s, u → u, a → a, l → l

c → k, ac → ak, c → kg, ac → ak, ct → kgs
```
Other Features

• Issues:
  – to learn: “economic” – “économique”
  – has ending mutation: “ic$” – “ique$”
  – requires a length-5 substring

• Solution:
  – Include all (arbitrary-length) substrings with aligned end characters, mismatching middles

• Also: Include NED as a feature
Learning Approach

• Support Vector Machine, linear kernel
  – optimize regularization parameter on dev. set
  – score pairs by positive distance from SVM hyperplane
Cognate Data Generation

• Is LCSR $\geq 0.58$ a good working definition of cognation? French-English Dictionary:
# System Development

| System                                      | Prec |
|---------------------------------------------|------|
| Klementiev-Roth (KR) $L\leq 2$              | 58.6 |
| KR $L\leq 2$ (normalized, boundary markers) | 62.9 |
| *phrases* $L\leq 2$                         | 61.0 |
| *phrases* $L\leq 3$                         | 65.1 |
| *phrases* $L\leq 3 + mismatches$            | 65.6 |
| *phrases* $L\leq 3 + mismatches + NED$      | 65.8 |

Table 2: Bitext French-English *development set* cognate identification 11-pt average precision (%).
## Example Most-Similar Words

| Greek-English – Dictionary | Spanish-English - Bitext |
|---------------------------|--------------------------|
| alkali:alkali              | agenda:agenda            |
| makaroni:macaroni*        | natural:natural          |
| adrenalini:adrenaline      | márgenes:margins         |
| flamingko:flamingo        | hormonal:hormonal        |
| spasmodikos:spasmodic      | radón:radon              |
| amvrosia:ambrosia         | higiénico:hygenic        |
Other Approaches

• Ristad & Yаниlos (1999)
  – stochastic transducer version of Edit Distance
  – can work with string pairs from different alphabets

• CRFs – learn to align as well as calculate similarity