Experiments
with
Normalized Compression Metric

Kiril Simov and Petya Osenova

Linguistic Modelling Laboratory

Bulgarian Academy of Sciences

(http://www.BulTreeBank.org)

BulDialects Project

5-6 December 2005

Tübingen, Germany

Plan of the Talk

- Similarity Metrics based on Compression
  (based on: Rudi Cilibrasi and Paul Vitanyi, *Clustering by Compression*, IEEE Trans. Information Theory, 51:4(2005) Also: http://www.cwi.nl/~paulv/papers/cluster.pdf (2003).)
- Experiments
- Conclusion
- Future Work

Feature-Based Similarity

- Task: Establishing of similarity between different data sets
- Each data set is characterized by a set of features and their values
- Different classifiers for definition of similarity
- Problem: definition of features, which features are important

Non-Feature Similarity

- The same task: Establishing of similarity between different data sets
- No features are specially compared
- Single similarity metric for all features
- Problem: the features that are important and play major role remain hidden in the data

Similarity Metric

- Metric: distance function $d(.,.)$ such that: $d(a,b)$?0; $d(a,b)=0$ iff $a=b$; $d(a,b)=d(b,a)$; $d(a,b)?d(a,c)+d(c,b)$ (triangle inequality)
- Density:

  For each object there are objects at different distances from it

- Normalization:

  The distance between two objects depends on the size of the objects. Distances are in the interval [0,1]

Kolmogorov Complexity

- For each file x, $k(x)$ (Kolmogorov complexity of x) is the length in bits of the ultimately compressed version of the file x (undecidable)
- Metric based on Kolmogorov complexity

  $k(x,y) = k(xy)$, where xy is the concatenation of x and y, is almost a metric:

  - $k(x,x) = k(xx)$? $k(x)$
  - $k(x,y) = k(y,x)$
  - $k(x,y)$? $\mathrm{k}(x,z) + k(z,y)$

Normalized Kolmogorov Metric

- A normalized Kolmogorov metric has to consider also Kolmogorov complexity of *x* and *y*
- We can see that

$$min(k(x),k(y)) \ ? \ k(x,y) \ ? \ k(x) + k(y)$$

$$0 \ ? \ k(x,y) - min(k(x),k(y)) \ ? \ k(x) + k(y) - min(k(x),k(y))$$

$$0 \ ? \ k(x,y) - min(k(x),k(y)) \ ? \ max(k(x),k(y))$$

$$0 \ ? \ (\ k(x,y) - min(k(x),k(y)) \ ) \ / \ max(k(x),k(y)) \ ? \ 1$$

Normalized Compression Distance

- Kolmogorov complexity is undecidable
- Thus, it can be only approximated by a real life compressor *c*
- Normalized compression distance *ncd*(.,.) is defined by

$$ncd(x,y) = (\ c(x,y) - min(c(x),c(y))\ )\ /\ max(c(x),c(y))$$

where *c*(*x*) is the size of the compressed file *x*

The properties of *ncd*(.,.) depends of the compressor *c*

Normal Compressor

- The compressor *c* is normal if it satisfies (asymptotically to the length of the files):
  - *Stream-basedness*: first *x*, then *y*
  - *Idempotency*: $c(xx) = c(x)$
  - *Symmetry*: $c(xy) = c(yx)$
  - *Distributivity*: $c(xy) + c(z)$? $c(xz) + c(yz)$
- If *c* is normal, then *ncd*(.,.) is a similarity metric

Problems with *ncd*(.,.)

- Real compressors are imperfect, thus *ncd*(.,.) is imperfect
- Good results can be obtained only for large data sets
- Each feature in the data set is a basis for a comparison
- Most compressors are byte-based, thus some intra-byte features can not be captured well

Real Compressors are Imperfect

- For a small data set the compression size depends on additional information like version number, etc
  - The compressed file could be bigger than the original file
- Some small reordering of the data does not play a role for the size of the compression
  - Series of ‘a b a b’ is treated the same as ‘a a b b’
- Substitution of one letter with another one could have no impact
- Cycles in the data are captured by the compressors

Large Dialectological Data Sets

- Ideally, large dialectological, naturally created data sets are necessary
- In practice, we can try to create such data by
  - Simulating ‘naturalness’
  - Hiding of features that are unimportant to the comparison of dialects
  - Encoding that allows direct comparison of the important features: p <-> b (no), p <-> p* (yes)

Generation of Dialectological Data Sets

- We decided to generate dialectological ‘texts’
- First we did some experiments with non-dialectological data in order to study the characteristics of the compressor. Results show:
  - The repetition of the lexical items has to be non cyclic
  - The features explication needs to be systematic
  - Linear order has to be the same for each site

Experiment Setup

- We have used the 36 words from the experiments of Petya in Groningen, transcribed in X-Sampa
- We have selected ten villages which are grouped in three clusters by the methods developed in Groningen:
  - [Alfatar, Kulina-voda]
  - [Babek, Malomir, Srem]
  - [Butovo, Bylgarsko-Slivovo, Hadjidimitrovo, Kozlovets, Tsarevets]

Corpus-Based Text Generation

The idea is the result to be as much as possible close to a natural text. We performed the following step:

- From a corpus of about 55 million words we deleted all word forms except for the 36 from the list
- Then we concatenated all remaining word forms in one document
- For each site we substituted the normal word forms with corresponding dialect word forms

Distances for Corpus-Based Text

v/v	Alfatar	Babek	Butovo	Bylgarsko -Slivovo	Hadjidi- mitrovo	Kozlo- vets	Kulina- voda	Malomir	Srem	Tsare- vets
Alfatar	0	0.958333	0.967278	0.967483	0.962608	0.967483	0.991503	0.95831	0.967673	0.967483
Babek	0.958333	0	0.989423	0.989575	0.987506	0.989575	0.99279	0.98481	0.983932	0.989575
Butovo	0.967278	0.989423	0	0.036648	0.62143	0.036529	0.973484	0.663445	0.507177	0.036529
Bylgarsko -Slivovo	0.967483	0.989575	0.036648	0	0.624508	0.002325	0.973821	0.662424	0.659798	0.002325
Hadji- dimitrovo	0.962608	0.987506	0.62143	0.624508	0	0.624917	0.969873	0.466019	0.758424	0.624917
Kozlovets	0.967483	0.989575	0.036529	0.002325	0.624917	0	0.973817	0.662382	0.506707	0.002202
Kulina- voda	0.991503	0.99279	0.973484	0.973821	0.969873	0.973817	0	0.97489	0.979109	0.972944
Malomir	0.95831	0.98481	0.663445	0.662424	0.466019	0.662382	0.97489	0	0.70567	0.660543
Srem	0.967673	0.983932	0.507177	0.659798	0.758424	0.506707	0.979109	0.70567	0	0.520216
Tsarevets	0.967483	0.989575	0.036529	0.002325	0.624917	0.002202	0.972944	0.660543	0.520216	0

Clusters According to Corpus-Based Text

- – [0.96 Kulina-voda]
- – [0.95 Alfatar]
- – [0.95 Babek]
- – [0.70 [0,46 Hadjidimitrovo, Malomir], Srem]
- – [0.03 Butovo, [0.003 Bylgarsko-Slivovo, Kozlovets, Tsarevets]]

Some Preliminary Analyses

- More frequent word forms play a bigger role: ???? – 106246 times vs. ??????? – 5 times from 230100 word forms
- The repetition of the word forms is not easily predictable thus close to natural text

Permutation-Based Text Generation

The idea is the result to be as much as possible with not predictable linear order. We performed the following step:

- All 36 words were manually segmented in meaningful segments: ['t_S','i','"r','E','S','a']
- Then for each site we did all permutation for each word and concatenated them:

["b,E,l,i]["b,E,i,l]["b,l,E,i]["b,l,i,E]["b,i,E,l]["b,i,l,E][E,"b,l,i]...

Distances for Permutation-Based Text

v/v	Alfatar	Babek	Butovo	Bylgarsko-Slivovo	Hadjidimitrovo	Kozlovets	Kulina-voda	Malomir	Srem	Tsarevets
Alfatar	0	0.714862	0.507658	0.483185	0.655673	0.531872	0.57006	0.432072	0.699153	0.479323
Babek	0.714862	0	0.658808	0.632702	0.572954	0.706679	0.551263	0.511125	0.288638	0.638389
Butovo	0.507658	0.658808	0	0.07827	0.361563	0.148523	0.723068	0.632968	0.717032	0.079008
Bylgarsko-Slivovo	0.483185	0.632702	0.07827	0	0.315238	0.099947	0.783802	0.661494	0.753367	0.014043
Hadjidimitrovo	0.655673	0.572954	0.361563	0.315238	0	0.360587	0.714916	0.668353	0.637938	0.259103
Kozlovets	0.531872	0.706679	0.148523	0.099947	0.360587	0	0.751512	0.746026	0.744859	0.058654
Kulina-voda	0.57006	0.551263	0.723068	0.783802	0.714916	0.751512	0	0.422748	0.588394	0.679138
Malomir	0.432072	0.511125	0.632968	0.661494	0.668353	0.746026	0.422748	0	0.578341	0.619165
Srem	0.699153	0.288638	0.717032	0.753367	0.637938	0.744859	0.588394	0.578341	0	0.64361
Tsarevets	0.479323	0.638389	0.079008	0.014043	0.259103	0.058654	0.679138	0.619165	0.64361	0

Clusters According to Permutation-Based Text

- [0.57 Kulina-voda, [0.43 Alfatar, Malomir]]
- [0.28 Babek, Srem]
- [0.25 Hadjidimitrovo, [0.07 Butovo, Bylgarsko-Slivovo, Kozlovets, Tsarevets]]

Conclusions

- Compression methods are feasible with generated data sets
- Two different measurements of the distance of dialects:
  - Presence of given features
  - Additionally distribution of the features

Future Work

- Evaluation with different compressors
- Better explication of the features
- Better text generation: more words and application of (sure) rules
- Implementation of the whole process of application of the method
- Comparison with other methods
- Expert validation