DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute
http://kt.jis.si

Efficient Multilingual Lemmatisation

Matjaž Juršič, Igor Mozetič, Nada Lavrač
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

What is lemmatisation?

process of determining a lemma of a given word

LEMMA
- canonical form of a word
- usually corresponds to a headword in a dictionary

MOTIVATION
- important step during pre-processing text for majority knowledge discovery methods

Word	Lemma
pišem	pisati
piše	pisati
pisali	pisati
pisali	pisalo
pisalom	pisalo

What are Ripple Down Rules - RDR?

- incremental knowledge acquisition methodology
- knowledge representation formalism

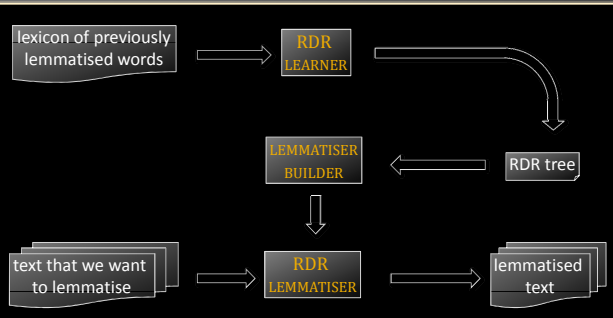
DATA STRUCTURE
- tree-like decision structure
- one node is one extended if-then rule

IF (bird) THEN (flies)
  EXCEPT IF (young bird) THEN (doesn't fly)
  ELSE IF (penguin) THEN (doesn't fly)
    EXCEPT IF (penguin in airplane) THEN (flies)
ELSE IF (airplane) THEN (flies)

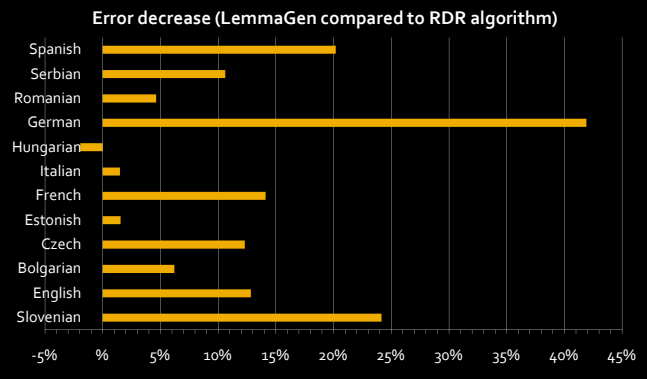
RDR on lemmatisation Domain

1 (if suffix = "", then ""->"") except
  1.1 (if suffix = "i", then "i"->"o") except
    1.1.1 (if suffix = "li", then "li"->"ti")
    1.1.2 (if suffix = "ni", then "ni"->"ti")
    1.1.3 (if suffix = "ti", then ""->"")
    1.1.4 (if suffix = "ši", then "ši"->"sati")
  1.2 (if suffix = "l", then "l"->"ti")
  1.3 (if suffix = "mo", then ""->"") except
    1.3.1 (if suffix = "šemo", then "šemo"->"sati")
    1.3.2 (if suffix = "šimo", then "šimo"->"sati")

LemmaGen Architecture



Evaluation results



	Language	Accuracy (%)								
		Known words (optimistic)			Random test words (realistic)			Unknown words (pessimistic)		
		RDR	L-Gen	Errors	RDR	L-Gen	Errors	RDR	L-Gen	Errors
MULTEXT- EAST	Slovenian	95,35	97,61	-48,6	92,59	94,38	-24,1	80,68	82,12	-7,5
	Serbian	94,36	97,86	-62,1	70,34	73,49	-10,6	64,26	65,85	-4,5
	Bulgarian	91,22	93,68	-28,0	74,52	76,10	-6,2	69,29	71,52	-7,2
	Czech	96,61	97,89	-37,8	92,77	93,66	-12,3	78,09	81,13	-13,9
	English	97,75	98,84	-48,3	92,05	93,07	-12,8	89,27	91,03	-16,4
	Estonian	86,81	89,51	-20,5	73,52	73,93	-1,6	66,69	66,54	0,5
	French	96,72	98,80	-63,5	91,78	92,94	-14,1	86,80	88,22	-10,8
	Hungarian	90,23	91,88	-16,9	74,82	74,33	2,0	72,73	72,86	-0,5
	Romanian	94,96	96,75	-35,6	78,16	79,17	-4,6	73,48	74,14	-2,5
MULTEXT	English	98,20	99,00	-44,5	93,29	94,14	-12,7	90,82	92,48	-18,1
	French	96,72	98,80	-63,5	91,79	92,95	-14,2	86,85	88,25	-10,7
	German	95,88	98,70	-68,5	95,06	97,13	-41,9	79,56	84,15	-22,4
	Italian	93,75	95,58	-29,2	85,87	86,08	-1,5	82,05	82,11	-0,3
	Spanish	99,10	99,48	-42,1	94,65	95,73	-20,1	94,32	95,45	-19,9

Results of This Work

SET OF UTILITIES FOR LEMMATISATION
- learning, building: (*LemLearn, LemBuild*)
- application: (*lemmatise*)
- evaluation: (*LemXval, LemTest, LemSplit, LemStat*)

OPEN SOURCE LEMMAGEN LIBRARY IN C++
- complete lemmatisation functionality for integration in more complex systems

LEMMATISERS FOR 12 EUROPEAN LANGUAGES
- pre-learned, pre-build and ready to use lemmatisers
- suitability estimation of RDR principle for each language

Further Work

IMPROVED LEMMATISERS
- Higher accuracy using better and larger lexicons
- New lexicons for new languages
- Learning algorithm improvements

ACCESSIBILITY OF DEVELOPED METHODS
- Integration in existing systems for text mining
- Web service for lemmatisation

POSSIBLE APPLICATION ON OTHER DOMAINS
- Transformation of serialized data between two similar forms that have different suffix

COMPARISON WITH OTHER LEMMATISATON METHODS