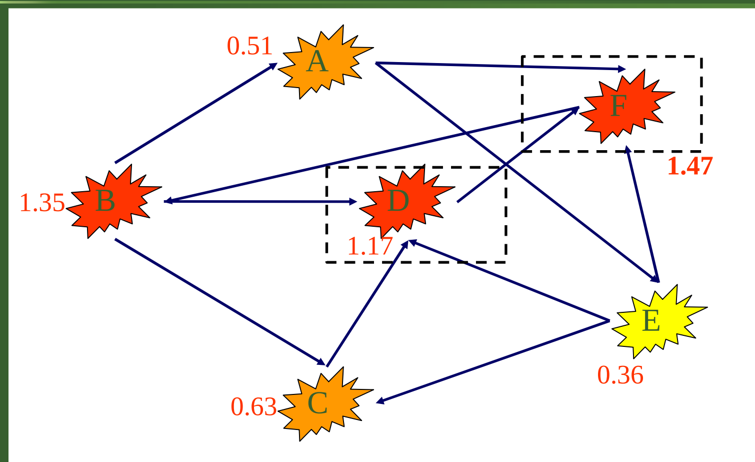
Automatic Keyword Extraction from Learning Objects

Kino Coursey and Rada Mihalcea

Our Experiments so Far…

- TextRank
  - Graph-based keyword extraction
- Wikifier
  - Algorithm based on the Wikipedia repository
- Combining the two methods
  - Intersection based on the “least common substring”
- All the evaluations carried out on the History course data from Phase I

TextRank: Random walk algorithms for natural language processing



Rada Mihalcea and Paul Tarau, "TextRank: Bringing Order into Texts" EMNLP 2004.

Random Walk Algorithms

- Usually applied on directed graphs
  - From a given vertex, the walker selects at random one of the out-edges
- Given G = (V,E) a directed graph with vertices V and edges E
  - In(Vi) = predecessors of Vi
  - Out(Vi) = successors of Vi

$$S(V_i) = (1-d) + d \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

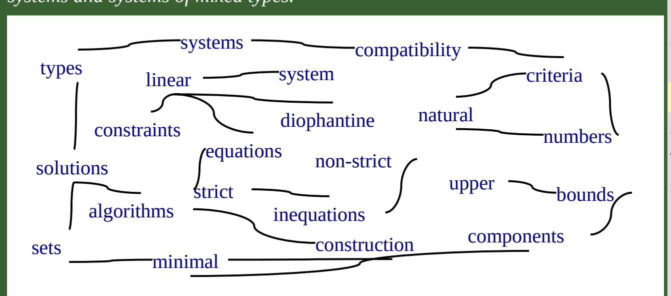
d – damping factor $\in [0,1]$ (usually 0.85)

TextRank for Keyword Extraction

- Store words in vertices
- Use co-occurrence to draw edges
- Rank graph vertices across the entire text
- Pick top N as keywords

An Example

Compatibility of systems of linear constraints over the set of natural numbers Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.



TextRank
numbers (1.46)
inequations (1.45)
linear (1.29)
diophantine (1.28)
upper (0.99)
bounds (0.99)
strict (0.77)

Frequency
systems (4)
types (4)
solutions (3)
minimal (3)
linear (2)
inequations (2)
algorithms (2)

Keywords by TextRank: *linear constraints, linear diophantine equations, natural numbers, non-strict inequations, strict inequations, upper bounds*
Keywords by human annotators: *linear constraints, linear diophantine equations, non-strict inequations, set of natural numbers, strict inequations, upper bounds*

Previous evaluation on INSPEC abstracts

- Evaluation:
  - 500 INSPEC abstracts
  - collection previously used in keyphrase extraction [Hulth 2003]
- Previous work
  - mostly supervised learning
  - [Hulth 2003]
  - training/development/test : 1000/500/500 abstracts

Method	Assigned Total	Assigned Mean	Correct Total	Correct Mean	Precision	Recall	F-measure
TextRank	6,784	13.7	2,116	4.2	**31.2**	43.1	**36.2**
Ngram with tag	7,815	15.6	1,973	3.9	25.2	**51.7**	33.9
NP-chunks with tag	4,788	9.6	1,421	2.8	29.7	37.2	33
Pattern with tag	7,012	14.0	1,523	3.1	21.7	39.9	28.1

Text Wikification

- Finding key terms in documents and link them to relevant encyclopedic information.

Lisbon

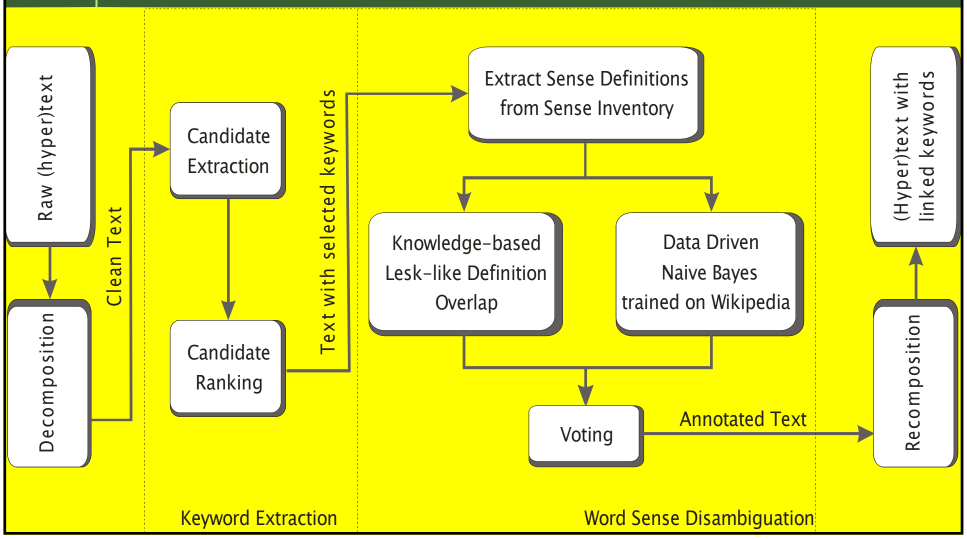
From Wikipedia, the free encyclopedia

For other uses, see Lisbon (disambiguation).

Lisbon (Portuguese: *Lisboa*, IPA: [liʒˈboɐ]) is the capital and largest city of Portugal. It is also the seat of the district of Lisbon and capital of the Lisbon region. Its municipality, which matches the city proper excluding the larger continuous conurbation, has a municipal population of 564,477[1] in 84.8 km², while the Lisbon Metropolitan Area in total has around 2.8 million inhabitants, and 3.34 million people live in the broader agglomeration of Lisbon Metropolitan Region (includes cities ranging from Leiria to Setúbal).[2] Due to its economic output, standard of living, and market size, the Grande Lisboa (Greater Lisbon) subregion is considered the second most important financial and economic center of the Iberian Peninsula. It is also the political center of the country, as seat of government and residence of the Head of State.

Rada Mihalcea and Andras Csomai, "Linking Documents to Encyclopedic Knowledge" CIKM 2007.

Wikification Pipeline



Keyword Extraction

- Semi-Controlled vocabulary
  - Wikipedia article titles and anchor texts (surface forms).
    - E.g. “USA”, “U.S.” = “United States of America”
  - 1.918.830 terms/phrases
  - Vocabulary is broad: “*the*” has 9 senses.
- Unsupervised keyword extraction
  - Tf * Idf
    - Wikipedia articles as document collection
  - Chi-squared independence of phrase and text
    - The degree to which it appeared more times than expected by chance
  - Keyphraseness: $$P(keyword \mid W) = \frac{count(D_{key})}{count(D_W)}$$

Previous Evaluation on Wikipedia

- 85 documents containing 7.286 links
- Extract *n* keywords, *n=6%* of number of words

	precision	recall	F-measure
Tf * Idf	41.91%	43.73%	42.82%
Chi-squared	41.44%	43.17%	42.30%
Keyphraseness	**53.37%**	**55.90%**	**54.63%**

Combining TextRank and Wikify!

- Using the strengths of both systems
  - TextRank focuses on estimating the attention given to terms in the text
  - Wikify focuses on keywords identified by a large number of people (Wikipedia)
  - Each gets a different set of interesting terms

The Violent Agreement Problem

- Two extractors with possibly different but complete segmentations of the same text.
  - "Mexican traveler" vs "Mexican"," traveler"
  - "Birth of Venus Sandros Botticelli" vs "The Birth of Venus", "Sandros Botticelli"
- TextRank gets extended noun-chunks while Wikify! gets common key phrases or object identifiers
- Need a principled way to find agreement
  - Intersection, Union, longest common substring

LCS : Longest Common Substring

- Problem: Given sequences x[1..m] and y[1..n], find a longest common subsequence of both.
- Example: x=BDABCBADAB and y=BDBCBABDAB,
  - BCB is a common substring and
  - BCBA and BDAB are two LCSs
- Common problem for aligning two DNA sequences
- Uses a dynamic programming method to find the longest common path through both strings
- In Subsequence (vs Substring) one allows gaps and is related to minimum Edit Distance

http://en.wikibooks.org/wiki/Algorithm_implementation/Strings/Longest_common_substring

http://en.wikipedia.org/wiki/Longest_common_substring_problem

LCS Example

- Applies to "Birth of Venus Sandros Botticelli" vs "The Birth of Venus", "Sandros Botticelli"
  - LCS("Birth of Venus Sandros Botticelli" ,"The Birth of Venus")="Birth of Venus"
  - LCS("Birth of Venus Sandros Botticelli","Sandros Botticelli")="Sandros Botticelli"
- Do a cross comparison for the output of both keyword sources keeping the longest match found for each
- Captures *coherent fragments* found by both

LCS Algorithm

L	S	Birth	of	Venus	Sandros	Botticelli
T	0	0	0	0	0	0
The	0	0	0	0	0	0
Birth	0	**1**	0	0	0	0
of	0	0	**2**	0	0	0
Venus	0	0	0	**3**	0	0

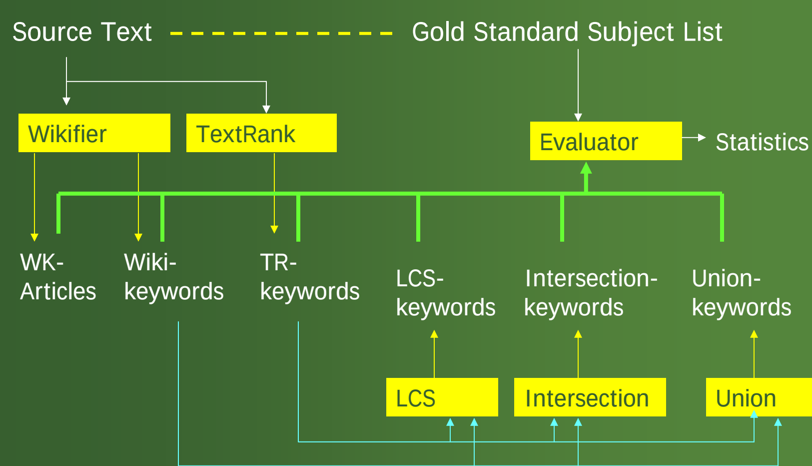
function LCSubstr(S[1..m], T[1..n])
L := **array**(0..m, 0..n)
z := 0 (length of longest match)
ret := {} (set of longest matches)
for i := 1..m
  for j := 1..n
    if S[i] = T[j] **then** L[i,j] := L[i-1,j-1] + 1 (the upper left diagonal)
    if L[i,j] > z **then** z := L[i,j] ret := {} (new longest found)
    if L[i,j] = z **then** ret := ret ∪ {S[i-z+1..i]} (an equal longest found)
return ret

Returns set of all matches of maximal length in one pass through the two strings

Intersection and Union

- Intersection
  - Create a list of words and phrases common to both list
- Union
  - Create a list of words and phrases on either list

The System



Evaluations on History LO

- Goal: Given the export of the text of the Learning Objects determine the performance of the various methods
  - Precision and recall for basic whole keyword extraction
  - Individual words in the text being correctly classified as being in the bag-of-gold keywords

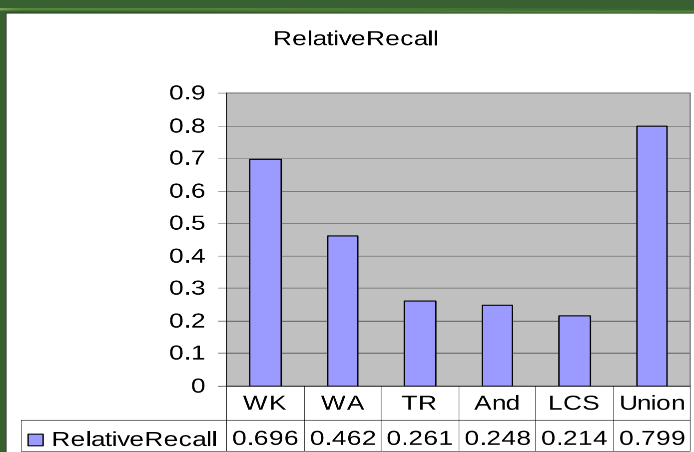
Gold Standard and its use

- Each collection of learning objects in a directory has a Dublin Core description file with subjects specified. These subjects are the gold standard.
  - Issue 1: One set of keyword for a set of files
  - Issue 2: The set of keywords may not have any direct reference in any of the text
- Each file assumes that the gold for it is the gold found in the appropriate Dublin file
- A pseudo-document is created consisting of all the text in the group to test the Dublin keywords against all the text the Dublin file covers

The Flash Card Problem

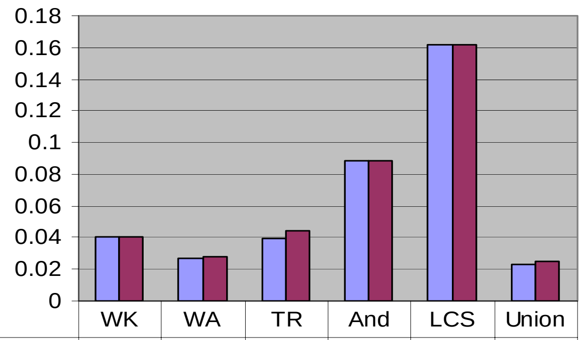
- Several html files are labeled “flash cards” and contain the following text : “the flash cards”
- Each card has different gold standard sets
- Each contains the same data
- “the flash cards” is not in the gold standard
- Same data + different specified outcome + no valid clues = ???
  - –http://lit.csci.unt.edu/~rada/Viewer/the%20flash%20card%205.htm.txt.html

Relative Recall



Relative recall = number of keywords identified out of gold standard keywords that appear in the text

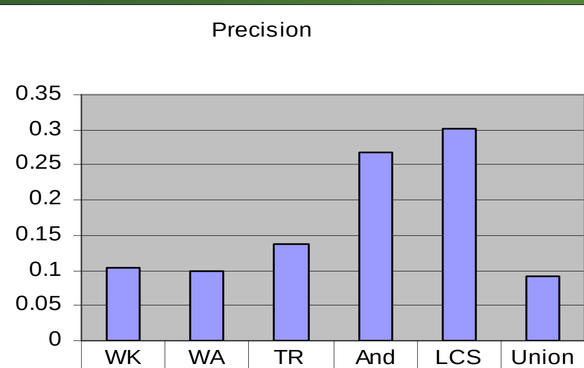
Keyphrase Precision



	WK	WA	TR	And	LCS	Union
keyPrecision1	0.0403	0.0267	0.0399	0.0889	0.162	0.0226
keyPrecision2	0.0401	0.028	0.0442	0.0888	0.1617	0.0246

keyPrecision1 =Gold keywords found in KeyList / Total Keylist size
keyPrecision2 =KeyList words found in Gold / Total KeyList size

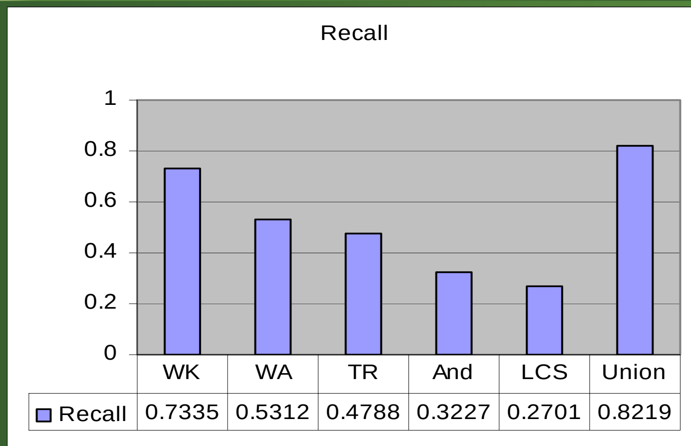
Word Level Precision



	WK	WA	TR	And	LCS	Union
Precision	0.1048	0.0993	0.1365	0.269	0.3008	0.0923

Word level precision = True positive / (True positive + false positive)
= # correct guesses / # guesses

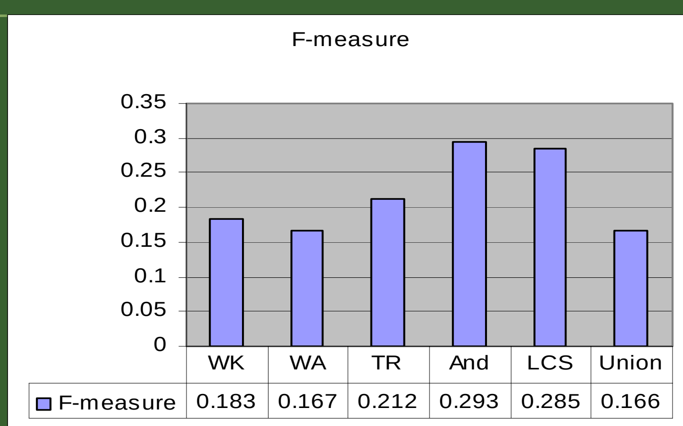
Word Level Recall



	WK	WA	TR	And	LCS	Union
Recall	0.7335	0.5312	0.4788	0.3227	0.2701	0.8219

Word level recall = True Positive / (True Positive + False negative)
= # correct guesses / # gold words

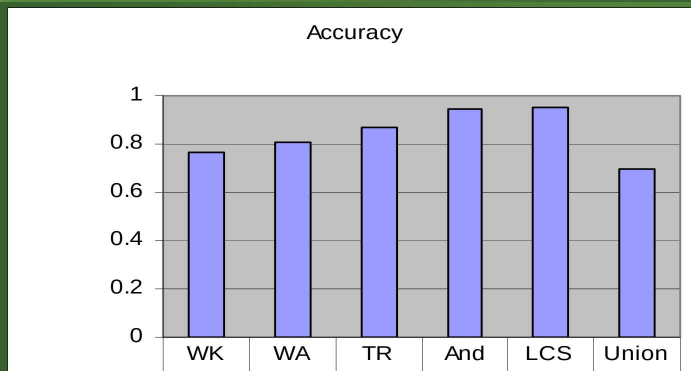
Word Level F-measure



	WK	WA	TR	And	LCS	Union
F-measure	0.183	0.167	0.212	0.293	0.285	0.166

Word level F-measure = (2* True positive) / (2*true positive + false positive + false negative)
= balance between recall and precision

Word Level Accuracy

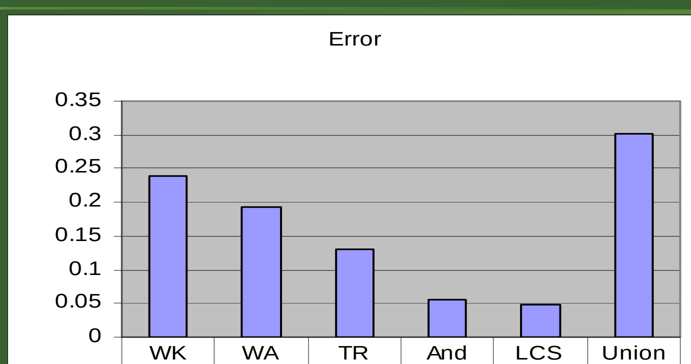


	WK	WA	TR	And	LCS	Union
Accuracy	0.7622	0.8076	0.8707	0.9434	0.9506	0.6992

Accuracy = (True positive + true negatives)/(all words)

= percentage of total word classifications correct

Word Level Error Rate



	WK	WA	TR	And	LCS	Union
Error	0.2378	0.1924	0.1293	0.0566	0.0494	0.3008

Error = 1 – Accuracy

= percentage of total possible misclassified

Learning Object Analysis

- Complete Analysis
  - http://lit.csci.unt.edu/~rada/Viewer
- Final Statistics
  - http://lit.csci.unt.edu/~rada/Viewer/SystemFinalSummary.html
- The Flash Card Problem
  - http://lit.csci.unt.edu/~rada/Viewer/the%20flash%20card%208.htm.txt.html
- Boston Tea Party
  - http://lit.csci.unt.edu/~rada/Viewer/boston_gazette.htm.txt.html
- Problems Facing the New Country
  - http://lit.csci.unt.edu/~rada/Viewer/07_problems_facing_new.htm.txt.html

Thoughts on Perfomance

- If you want high recall : Wikifier
  - Relative recall : 69%
  - Low precision : 4%
- If you are interested in balance: LCS
  - Recall: 21%
  - Precision: 16%

Questions, Thoughts ...

- Gold standard is not really “gold”
- Should we run a separate evaluation with users?
- What will be the end use of the automatic KE?
  - Emphasis on recall vs. emphasis on precision
- ???

Next Step ...

- Explore using the WikiArticles and performing WikiRank on them to get related articles and their keywords
  - A way to find higher level concepts not mentioned like “Revolutionary War” or “United States History” from a set of battles.