DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Experiments to Evaluate a Statistical Stemming Algorithm

M. Bacchin, N. Ferro, M. Melucci

Information Management Systems Research Group

Department of Information Engineering
University of Padua – Italy

michela.bacchin@unipd.it

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Stemming Process

- To design a stemming algorithm it is possible to follow at least two approaches:
  - Based on a-priori linguistic knowledge
  - Based on statistical methods which infer knowledge

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

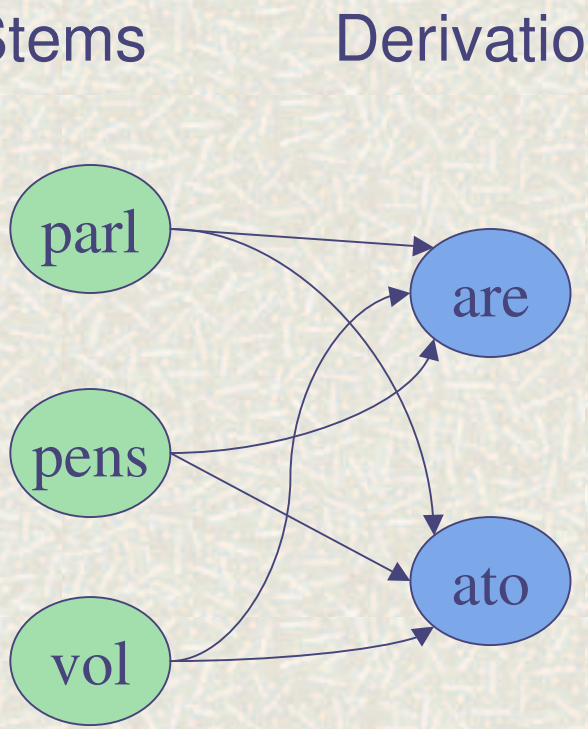
SPLIT: Key Concepts

- Suffix stripping paradigm.
- We build a collection of substrings extracted from words.
- We use a graph notation to represent the collection of substrings: nodes are substrings and an edge exists between 2 nodes only if these 2 substrings form a word.
- Mutual Reinforcing Relationship among prefixes which are stems and suffixes which are derivations.

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Words Graph Notation

Word	Stem	Derivation
parlare	parl	are
parlato	parl	ato
pensare	pens	are
pensato	pens	ato
volare	vol	are
volato	vol	ato



DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

The Probabilistic Approach

- We are interested in looking for the prefix x^* such that:

$$x^* = \arg\max_x \; P(x \in S \mid w \in W) = \arg\max_x \frac{P(w \in W \mid x \in S) \cdot P(x \in S)}{P(w \in W)}$$

  - The first term is estimated by the reciprocal of the number of words starting by the substring x
  - The second term is estimated using an iterative algorithm which discloses the mutual reinforcing relationship between stems and derivations (HITS)

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Disclosing Mutual Reinforcing Relationship

- HITS (Hyperlink Induced Topic Search) was originally proposed by J. Kleinberg to discover authoritative web pages.
- In this context, we assign each substring z two scores:

$$s_z^n = \sum_{\forall\, x \text{ prefix of } z} p_x^{n-1} \qquad p_z^n = \sum_{\forall\, y \text{ suffix of } z} s_y^n$$

- We estimated $P(x \in S)$ by the prefix score p_x

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Experiments to Evaluate SPLIT

- SPLIT can perform as effectively as an algorithm developed on the basis of a-priori linguistic knowledge?
- We developed a prototype IR system, called IRON. It is based on the top of an open-source java library, called LUCENE. It implements a vector-spacel model and a *tf·idf* weighting scheme.
- The stemming algorithm was implemented with a set of tools called SPLIT, which carries out the HITS estimation of the probabilities we are interested in.

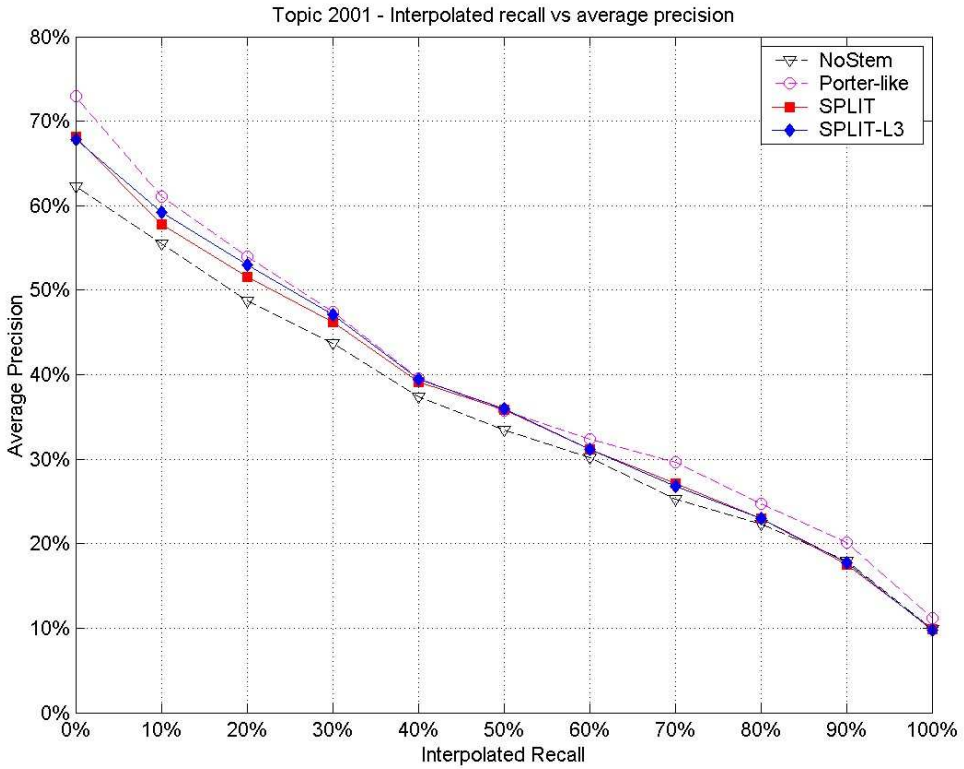
DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Experiments: Runs

- We compared the performances of IRON changing only the stemming algorithm for different runs, all other things being equal.
- We tested four different stemming algorithms:
  - NoStem: No stemming algorithm was applied.
  - Porter-like: An algorithm for the Italian language which applies a list of rules based on a-priori linguistic knowledge.
  - SPLIT: our statistical and graph-based stemming algorithm.
  - SPLIT-L3: the previous algorithm with a little ignition of linguistic knowledge (heuristic rule forcing the stem length to be at least 3).

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Experiments: 2001 Results

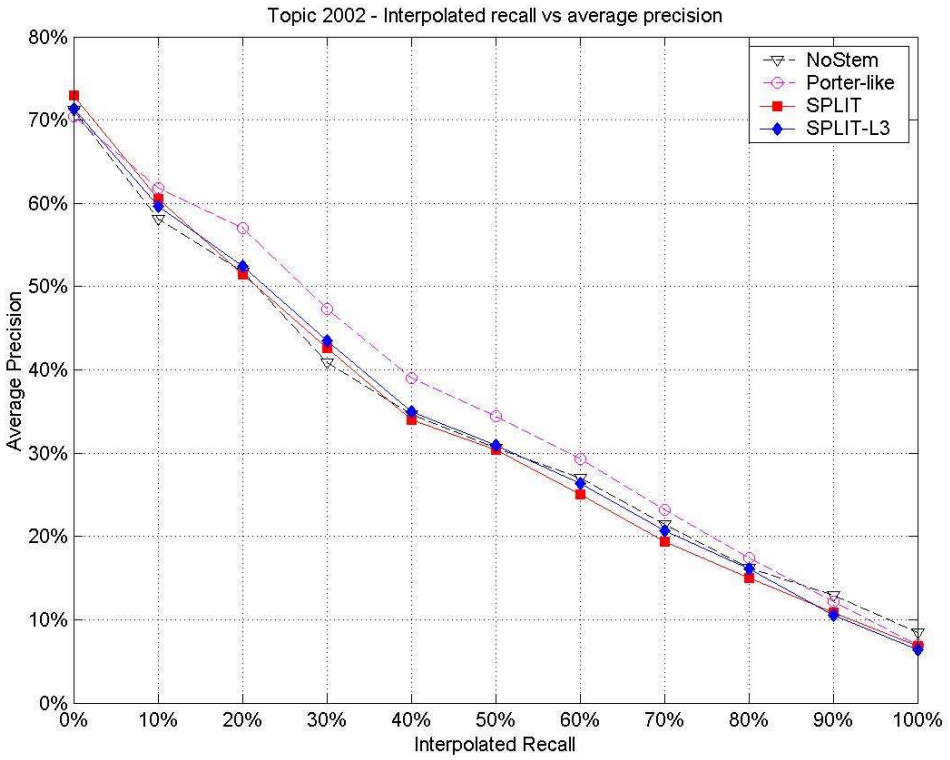


- For Avg-Prec e R-Prec all four methods are statistically equivalent.
- For Precision computed at 10, 20, 30 documents cut off values, stemming improves the performances, and SPLIT performs as effectively as Porter-like.

Algorithm	Avg-Prec
NoStem	0.3387
Porter	0.3753
SPLIT	0.3519
SPLIT-L3	0.3589

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Experiments: 2002 Results



- For R-Prec all four methods are again equivalent. For Avg-Prec there is a moderate statistical indication that Porter algorithm performs better than SPLIT.
- For Precision computed at 10, 20, 30 documents cut off values, all the methods are comparable.

Algorithm	Avg-Prec
NoStem	0.3193
Porter	0.3419
SPLIT	0.3173
SPLIT-L3	0.3200

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Conclusions and Future Work

- Objective: to investigate a stemming algorithm based on a link analysis procedure.
- The results are encouraging because the effectiveness level of SPLIT is comparable at least to that of an algorithm based on a-priori linguistic knowledge.
- Future work:
  - Further experiments with other languages, in order to test if it is a language-independent stemming algorithm.
  - To improve the probabilistic decision criterion.
  - To use a weighted graph model.

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Experiments to Evaluate a Statistical Stemming Algorithm

M. Bacchin, N. Ferro, M. Melucci

Information Management Systems Research Group

Department of Information Engineering
University of Padua – Italy

Thank you