Supervised Distributional Semantic Relatedness

Alistair Kennedy[1] Stan Szpakowicz[1,2]

School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada

Institute of Computer Science
Polish Academy of Sciences, Warsaw, Poland

TSD, 2012

Introduction

- Distributional measures of semantic relatedness (MSRs) determine word similarity based on how often word pairs appear in the same contexts.
- Often some measure of association is used to measure dependency between a word and a context.
- This is essentially an unsupervised process
  - How can this be made into a supervised process?
  - Can we use known sets of synonyms to enhance distributional MSRs?
- We build on earlier work from: [Kennedy and Szpakowicz(2011)].

Semantic Relatedness

- Semantic Relatedness
  - “cat” & “feline” – very similar
  - “cat” & “animal” – definitely related
  - “cat” & “hairdryer” – very little in common
  - “cat” & “math” – nothing in common really
- What is semantic relatedness used for?
  - Lexicon/Thesaurus construction, summarization, word sense disambiguation, etc.
- Common methods for measuring semantic relatedness
  - Resource based measures
  - Corpus based measures
  - Hybrid measures

Unsupervised MSRs

- Build a word-context matrix
  - Count how often each word appears in each context
- Most measures have two main parts
  - Re-weight the matrix
    - Association between words and contexts
    - PMI, LSA, etc.
  - Distance between words
    - Cosine similarity
- Motivation
  - Want to customize a measure for adding words to *Rogets Thesaurus*

Supervised MSRs

- Hypothesis: some contexts tend to be better indicators of synonymy than others
- Learn weights using known pairs of synonyms
  - Measure association between the words in a context and known synonyms
  - Use existing resources like *Rogets Thesaurus*, or *WordNet* as training data
- How to evaluate these measures?
  - Since enhancing *Roget's Thesaurus* is the goal, we use *Roget's* for evaluation.
  - Identify words that are in the same Head.
    - 1000 broad categories, e.g. *existence, nonexistence, beginning, end, etc.*

Supervised MSRs (2)

- Training Data examples
  - `omnipotence, omniscience, omnipresence`
  - `brotherhood, sisterhood`
  - `dial, sundial, gnomon, pendulum, hourglass`
  - `filling, stuffing, wadding, padding`
- Combined System
  - Combines best supervise and unsupervised MSRs.
  - Parameters are determined during a tuning phase.
  - Re-weights the word-context matrix twice: first using supervised re-weighting, then using unsupervised re-weighting

Our Experiments

- Build word-context matrix
  - Use *Minipar* dependency parser [Lin(1998)]
- Tuning Phase:
  - Evaluate six measures of association for *word-context* matrix re-weighting.
    - PMI, Z-score, T-score, Dice, Log Likelihood & χ^2
  - Evaluate two kinds of supervision
    - Training at the context level and relation level
- Testing Phase:
  - Evaluate each MSR on nouns, verbs and adjectives.
  - Explore different sources of training data:
    - *WordNet* and the 1911 and 1987 editions of *Roget's Thesaurus*.
  - Combine best supervised and unsupervised systems into one combined system

Building a Word-context Matrix

- Parse Wikipedia with *Minipar*
  - fin C:i:V settle
  - settle V:s:N ignorance
  - settle V:mod-before:A never
  - settle V:subj:N ignorance
  - settle V:obj:N question
  - question N:det:Det a
- "Ignorance never settles a question" Disraeli
- produces over 900 million dependency triples $\langle w, r, w' \rangle$
  - $\langle time, conj, motion \rangle$
  - time – $\langle conj, motion \rangle$ & motion – $\langle time, conj \rangle$
- Word w has context $\langle r, w' \rangle$ and word w' has context $\langle w, r \rangle$
- Construct matrix from counts of words and contexts

Re-weighting the Word-context Matrix

Observed and Expected Values

$$\begin{array}{c} \\ x \in X \\ x \notin X \end{array} \begin{array}{c} \begin{array}{cc} y \in Y & y \notin Y \end{array} \\ \begin{bmatrix} O_{0,0} & O_{0,1} \\ O_{1,0} & O_{1,1} \end{bmatrix} \end{array} \Rightarrow \begin{bmatrix} E_{0,0} & E_{0,1} \\ E_{1,0} & E_{1,1} \end{bmatrix}$$

$$E_{i,j} = \frac{\sum_y O_{i,y} \quad \sum_x O_{x,j}}{\sum_{x,y} O_{x,y}}$$

- Calculate association with one of the following measures:

$$Dice = \frac{2 * O_{0,0}}{\sum_j O_{0,j} + \sum_i O_{i,0}} \qquad PMI = \log \frac{O_{0,0}}{E_{0,0}}$$

$$Z\text{-}score = \frac{O_{0,0} - E_{0,0}}{\sqrt{E_{0,0}}} \qquad T\text{-}score = \frac{O_{0,0} - E_{0,0}}{\sqrt{O_{0,0}}}$$

$$\chi^2 = \sum_{i,j} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \qquad LL = 2 \sum_{i,j} O_{i,j} \log \frac{O_{i,j}}{E_{i,j}}$$

- This notation is borrowed from [Evert(2004)].

Measuring Association Unsupervised

- $O_{0,0}$ [True Positive] $[x \in X \wedge y \in Y]$:
  w_i is found in context c_j;
- $O_{0,1}$ [False Negative] $[x \in X \wedge y \notin Y]$:
  w_i is found in a context other than c_j;
- $O_{1,0}$ [False Positive] $[x \notin X \wedge y \in Y]$:
  a word other than w_i is found in context c_j;
- $O_{1,1}$ [True Negative] $[x \notin X \wedge y \notin Y]$:
  a word other than w_i is found in a context other than c_j.

Measuring Association – Unsupervised (2)

- For each context c and word w
- E.g. $c = \langle play, obj \rangle$, $w = hockey$
  - TP number of times hockey appears in $\langle play, obj \rangle$
  - FN number of times hockey appears in other contexts e.g. $\langle watch, obj \rangle$
  - FP number of times other words appears in $\langle play, obj \rangle$ e.g. football
  - TN number of times other words appear in other contexts e.g. $\langle watch, obj \rangle$ & football

Measuring Association – Supervised

- $O_{0,0}$ [True Positive] $[x \in X \wedge y \in Y]$:
  $\langle w_i, w_j \rangle$ are synonyms and both appear in c_k;
- $O_{0,1}$ [False Negative] $[x \in X \wedge y \notin Y]$:
  $\langle w_i, w_j \rangle$ are synonyms and only one appears in c_k;
- $O_{1,0}$ [False Positive] $[x \notin X \wedge y \in Y]$:
  $\langle w_i, w_j \rangle$ are not synonyms and both appear in c_k;
- $O_{1,1}$ [True Negative] $[x \notin X \wedge y \notin Y]$:
  $\langle w_i, w_j \rangle$ are not synonyms and only one appears in c_k.

Measuring Association – Supervised (2)

- For each context c
- E.g. $c = \langle play, obj \rangle$
  - Found in c: soccer, football, hockey
  - Not found with c: ice_hockey, airplane
- Count true positives, false positives, etc.
  - TP: soccer & football
  - FN: hockey & ice_hockey
  - FP: soccer & hockey
  - TN: hockey & airplane

Measuring Association – Supervised (3)

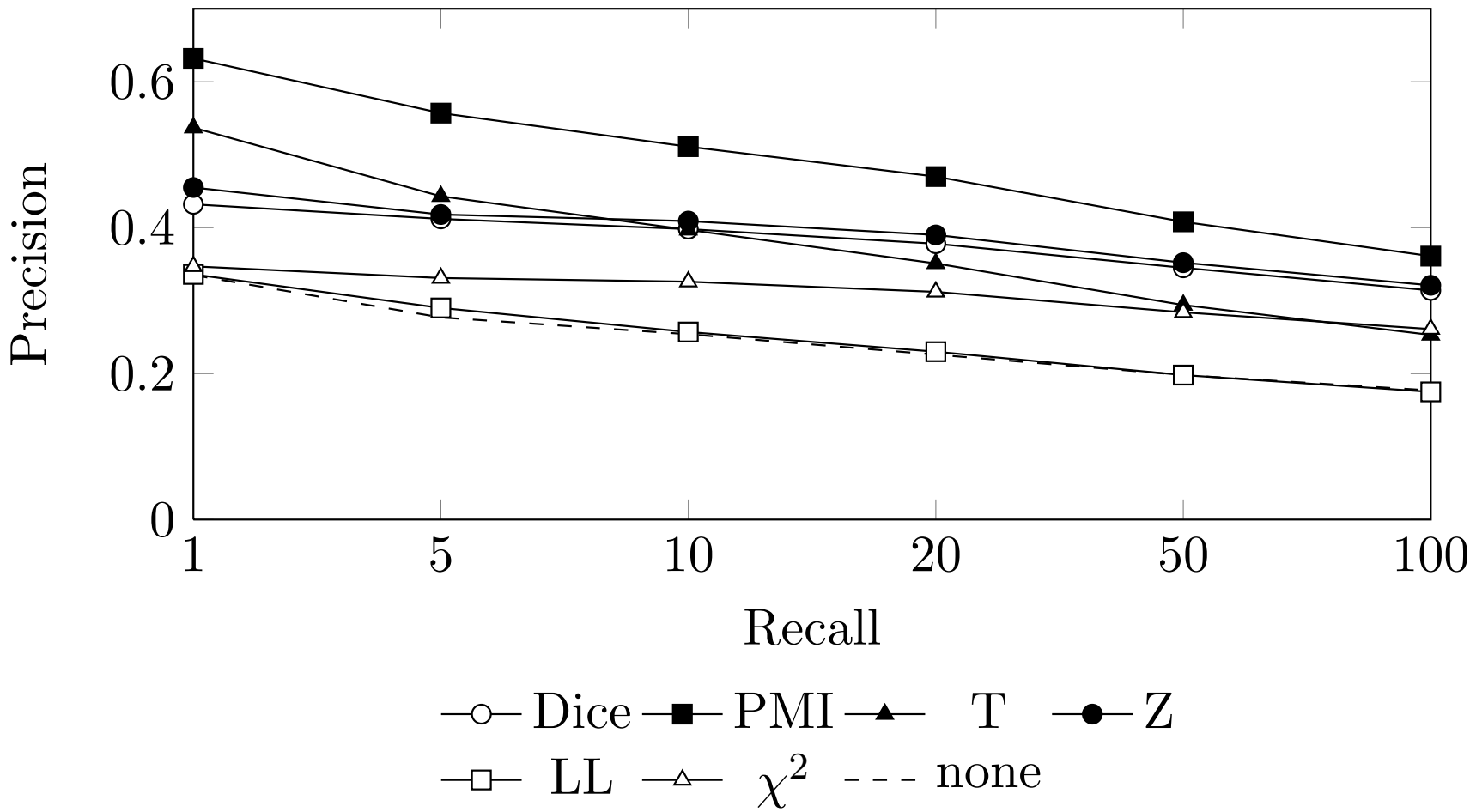
- Gives a unique weight for every context
- Can be altered to give a unique weight to every syntactic relation
  - Sum scores for each context sharing a common syntactic relation
  - Re-weighting at the *context-level* or *relation-level*
- Found best results when combining supervised and unsupervised methods
  - Re-weight word-context matrix with supervised re-weighting, then again using unsupervised re-weighting
  - Found to be best in [Kennedy and Szpakowicz(2011)].

Evaluation

- Identify closest words in *Roget's Thesaurus* 1987 edition.
  - How many nearest neighbours are found in the same Head at recall points of 1, 5, 10, 20, 50 and 100.
  - e.g. neighbours of "psychology" and their similarity scores:
    - sociology (0.720), anthropology (0.707), linguistics (0.582), economics (0.572)
- Tuning Phase
  - Determine best parameters: measure of association and training type
  - Use these best parameters in build a combined model
- Tuning and evaluation sets consist of 1000 nouns and 600 verbs & adjectives
  - All words from these data sets are excluded from the training data.

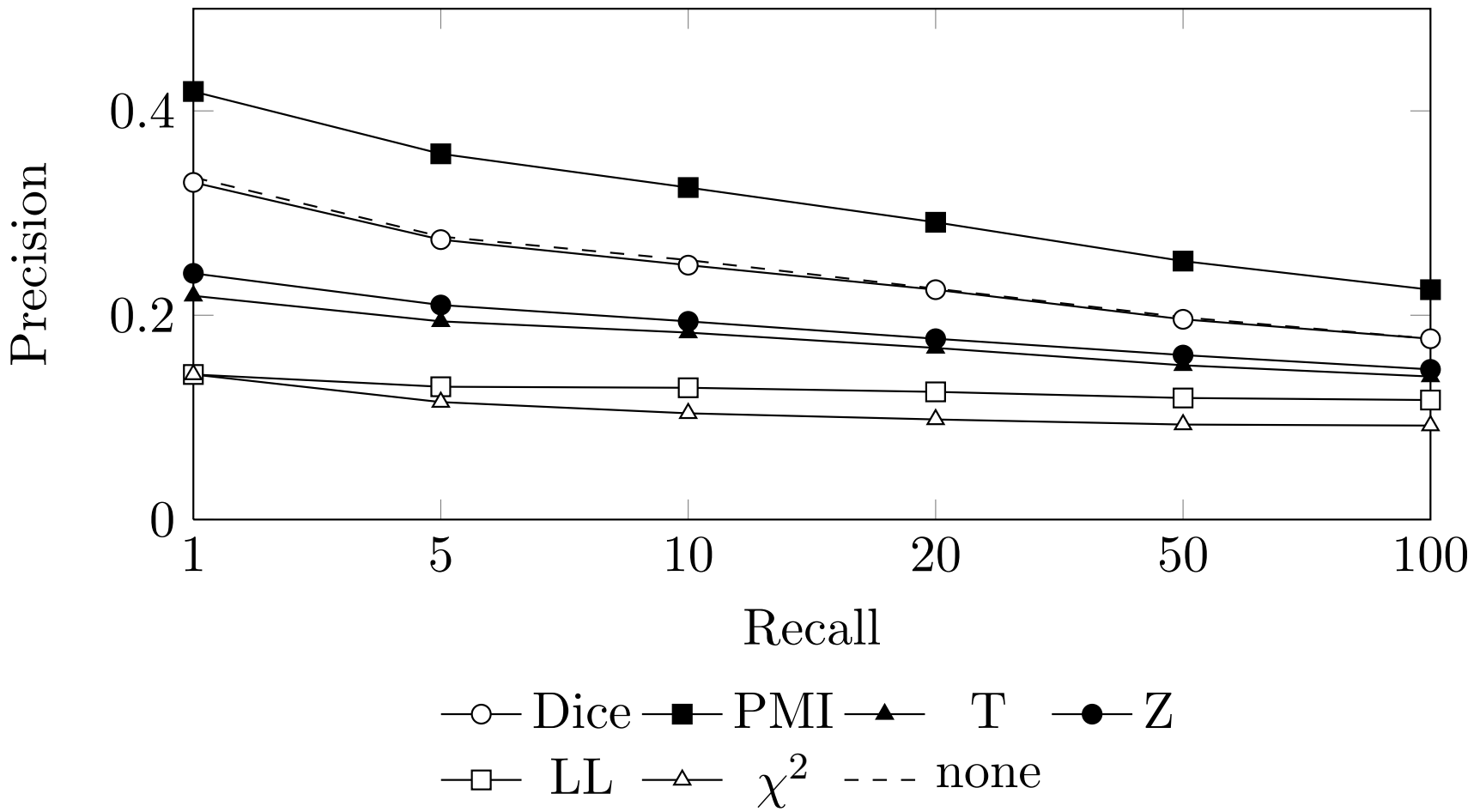
Tuning – Unsupervised Re-weighting

(Nouns)



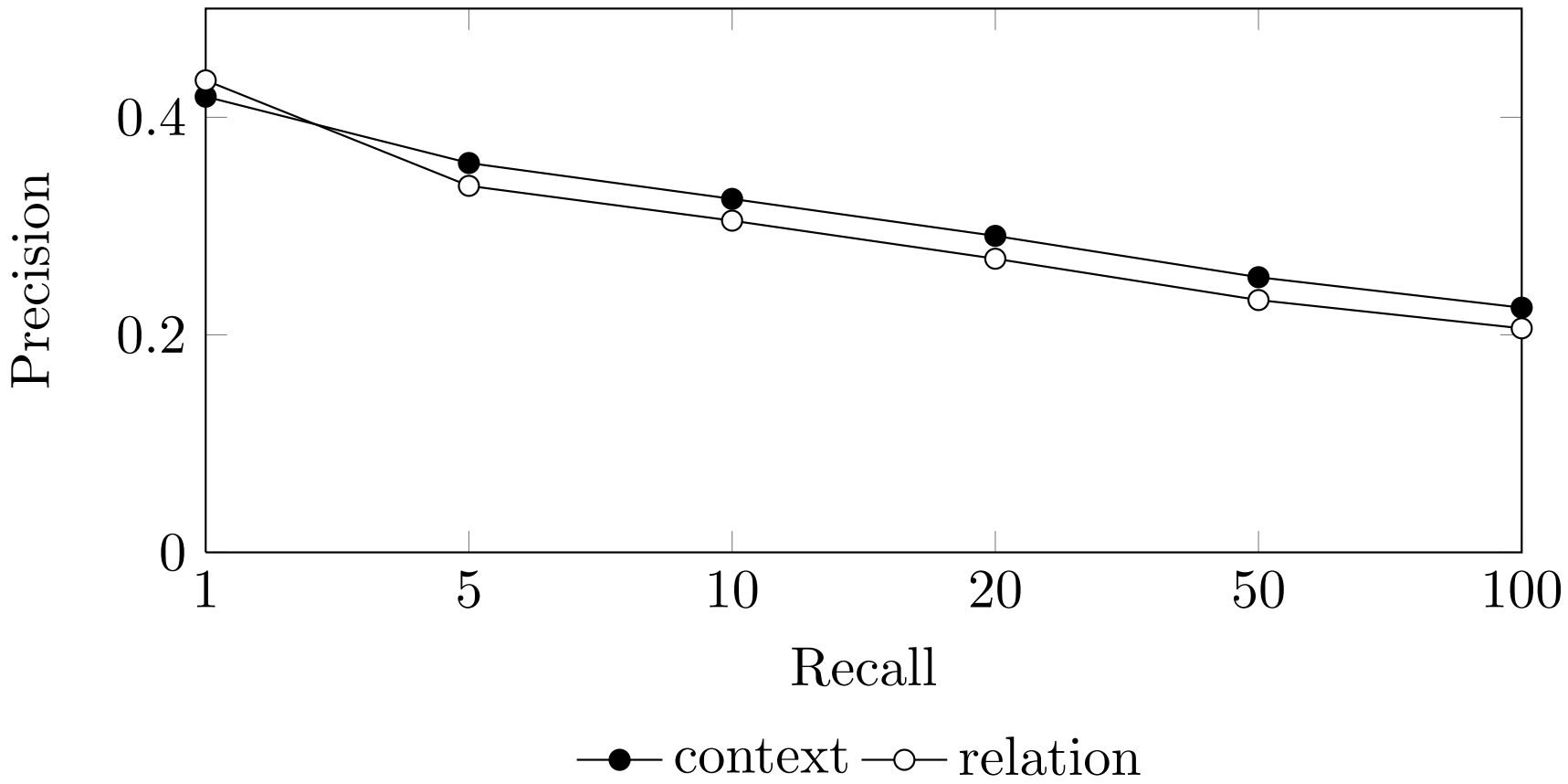
Tuning – Supervised Re-weighting

(Roget's 1911, Nouns)



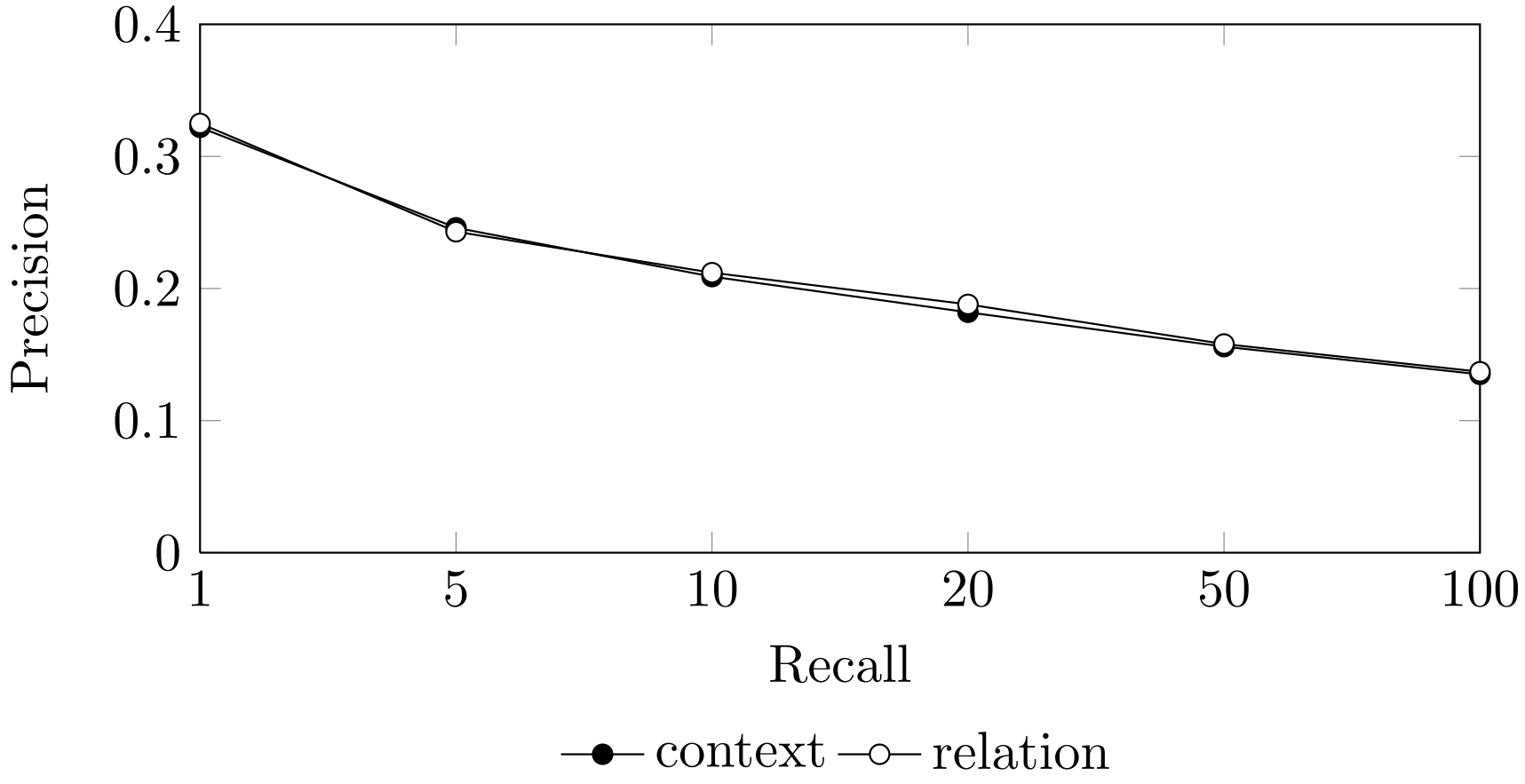
Tuning –Supervised Training type for Nouns and Verbs

(Roget's 1911)



Tuning – Supervised Train for Adjective

(Roget's 1911)



Final Evaluation

- Results from the tuning phase:
  - PMI proved to be the best measure of association
  - *context-level* re-weighting worked best for nouns and verbs
  - *relation-level* re-weighting worked best for adjectives
- Create combined system with these parameters
- Compare against two baselines
  - Low baseline – unweighted matrix
  - High baseline – unsupervised PMI-weighted matrix

Results for Nouns

Weight	Top 1	Top 5	Top 10	Top 20	Top 50	Top 100
Low	0.376	0.296	0.262	0.239	0.207	0.186
High	0.645	0.579	0.537	0.490	0.423	0.374
Combined-1911	0.659	**0.588**	**0.548**	**0.501**	**0.431**	**0.382**
Combined-1987	0.651	0.584	**0.549**	**0.501**	**0.430**	**0.381**
Combined-WN	0.654	0.586	0.541	**0.495**	**0.430**	**0.380**

Results for Verbs

Weight	Top 1	Top 5	Top 10	Top 20	Top 50	Top 100
Low	0.398	0.331	0.318	0.299	0.276	0.256
High	0.582	0.526	0.487	0.444	0.396	0.357
Combined-1911	0.605	0.533	**0.500**	**0.455**	**0.401**	**0.362**
Combined-1987	0.588	**0.537**	**0.499**	**0.453**	0.399	**0.360**
Combined-WN	0.587	0.531	**0.495**	**0.451**	0.395	0.356

Results for Adjectives

Weight	Top 1	Top 5	Top 10	Top 20	Top 50	Top 100
Low	0.317	0.259	0.224	0.205	0.163	0.139
High	0.600	0.480	0.431	0.368	0.295	0.247
Combined-1911	0.602	0.484	0.431	0.368	0.296	0.247
Combined-1987	0.603	0.483	0.431	0.367	0.296	0.247
Combined-WN	0.595	0.483	0.430	0.368	0.296	0.247

In Summary

- Best results came by mixing supervised and unsupervised re-weighting
- PMI would appear to be the best measure of association for building MSRs
- Found significant Improvement on Nouns and Verbs
- For Nouns and verbs out of 12 recall points
  - 1911 *Roget's* significantly improves 9
  - 1987 *Roget's* significantly improves 8
  - *WordNet* significantly improves 5
- No improvement for Adjectives
- Not entirely surpassing that *Roget's* performs better than *WordNet* as it is the source of training and testing data
  - That said, not clear either that Semicolon Groups should help train an MSR to find words in the same Head.

Bibliography

Stefan Evert.
The statistics of word cooccurrences: word pairs and collocations.
PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2004.

Alistair Kennedy and Stan Szpakowicz.
A supervised method of feature weighting for measuring semantic relatedness.
In *Proceedings of Canadian AI 2011*, pages 222–233, Ottawa, Ontario, Canada, 2011. Springer.

Dekang Lin.
Dependency-based evaluation of MINIPAR.
In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, 1998.