Erasmus Universiteit Rotterdam

Linguistic Graph Similarity for News Sentence Searching

Kim Schouten & Flavius Frasincar
schouten@ese.eur.nl *frasincar@ese.eur.nl*

Web News Sentence Searching Using Linguistic Graph Similarity, Kim Schouten and Flavius Frasincar. In *Proceedings of the 12th International Baltic Conference on Databases and Information Systems (DB&IS 2016)*, pages 319-333, Springer, 2016

COMMIT/ infinity

erasmus studio

ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus Universiteit Rotterdam

Problem

- Most text search methods are word-based
- Often, the context is lost for the sake of simplicity
- However, the meaning of a word is defined by both word and context
- How can we include context information of words into the search algorithm?
- Can we not search by sentence instead of words, and retrieve sentences with similar meaning?

ERASMUS UNIVERSITEIT ROTTERDAM

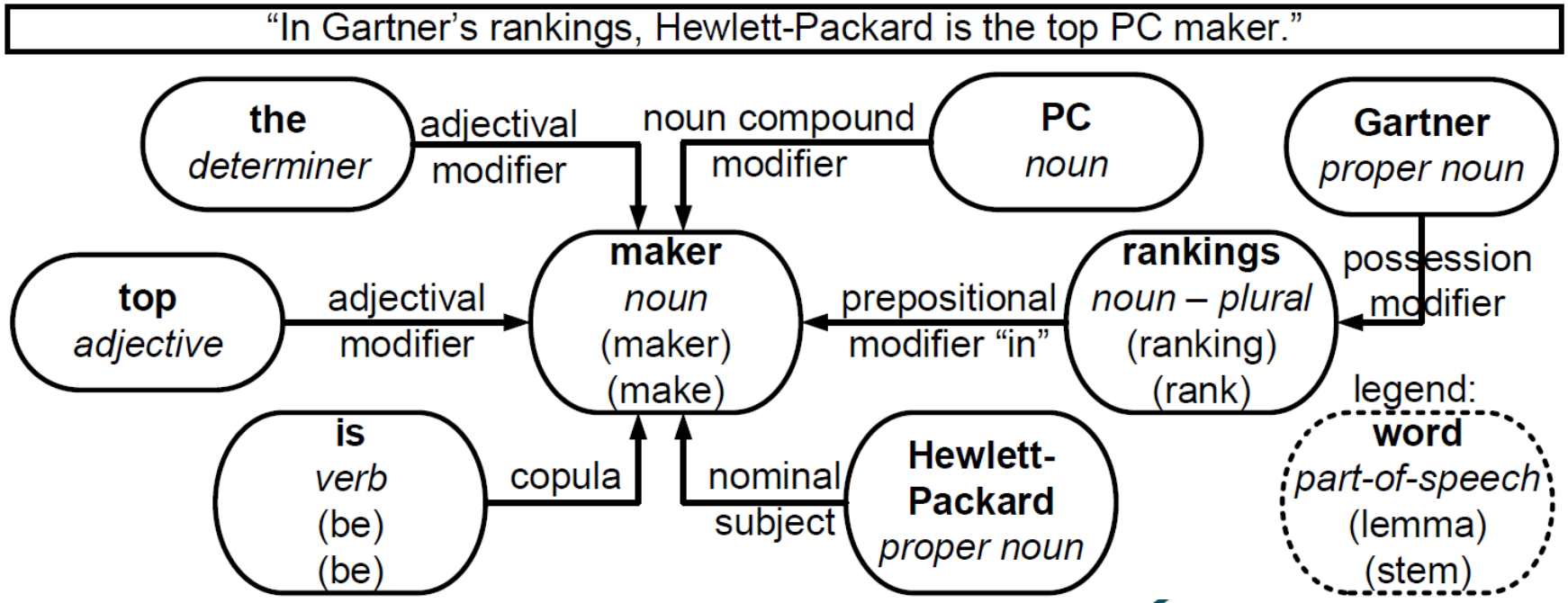
Erasmus Universiteit Rotterdam

Graph-based Approach

- Grammatically parsing a sentence yields a graph
  - Words are the nodes
  - Grammatical relations between words are the edges
- Set of relations of a word can then be used as context
- NLP pipeline transforms both query and news sentences into graphs

ERASMUS UNIVERSITEIT ROTTERDAM

Graph representation of sentence



Erasmus Universiteit Rotterdam

Graph comparison

- Problem is similar to graph isomorphism
- But *partial* similarity makes it much harder
  - Nodes may be missing on either side
  - Nodes may be only partially similar (pc <> workstation)
  - Relation labels may be different for similar nodes
- Hence, output is not binary but a real-valued similarity score

ERASMUS UNIVERSITEIT ROTTERDAM

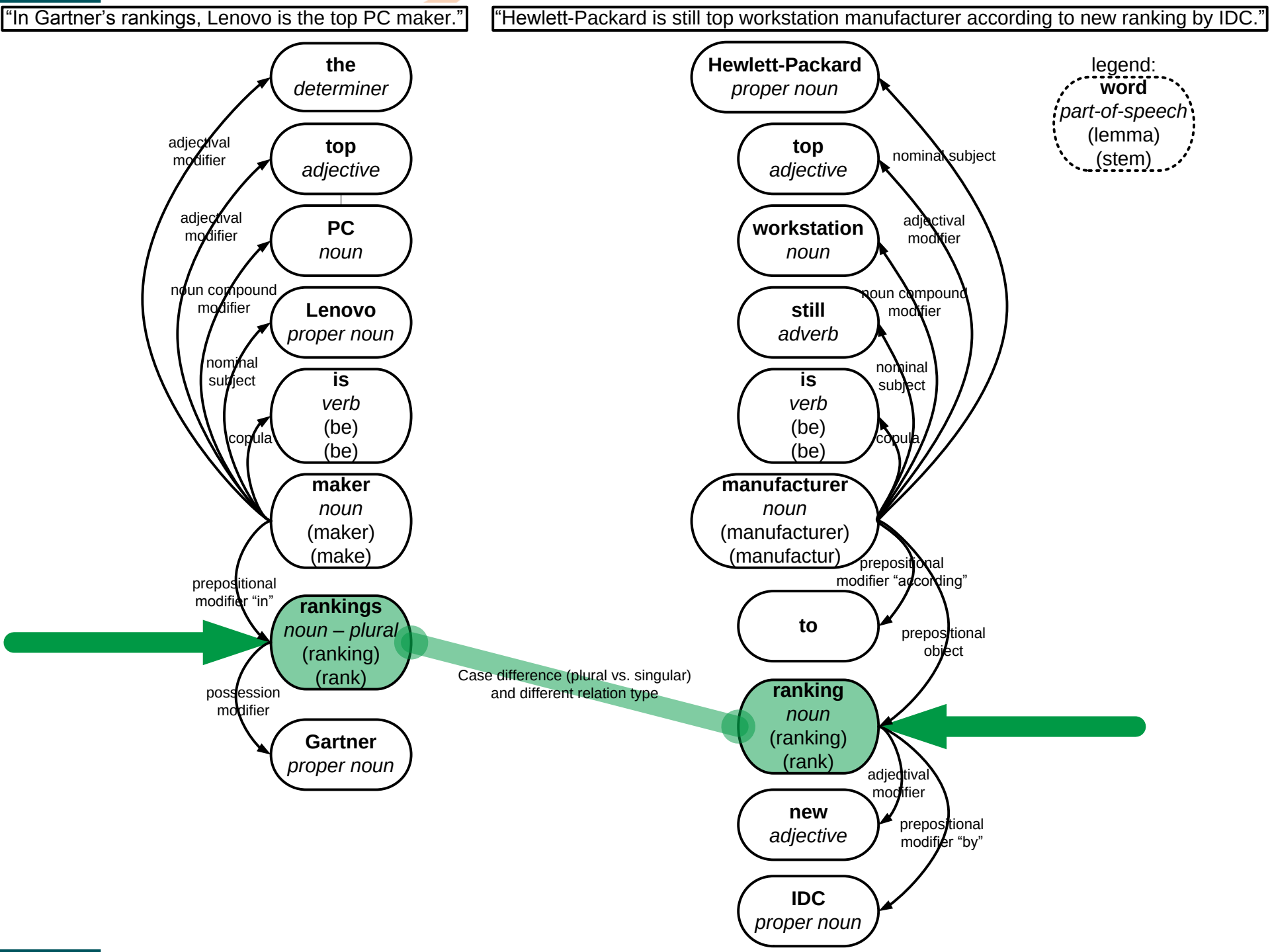
Graph comparison

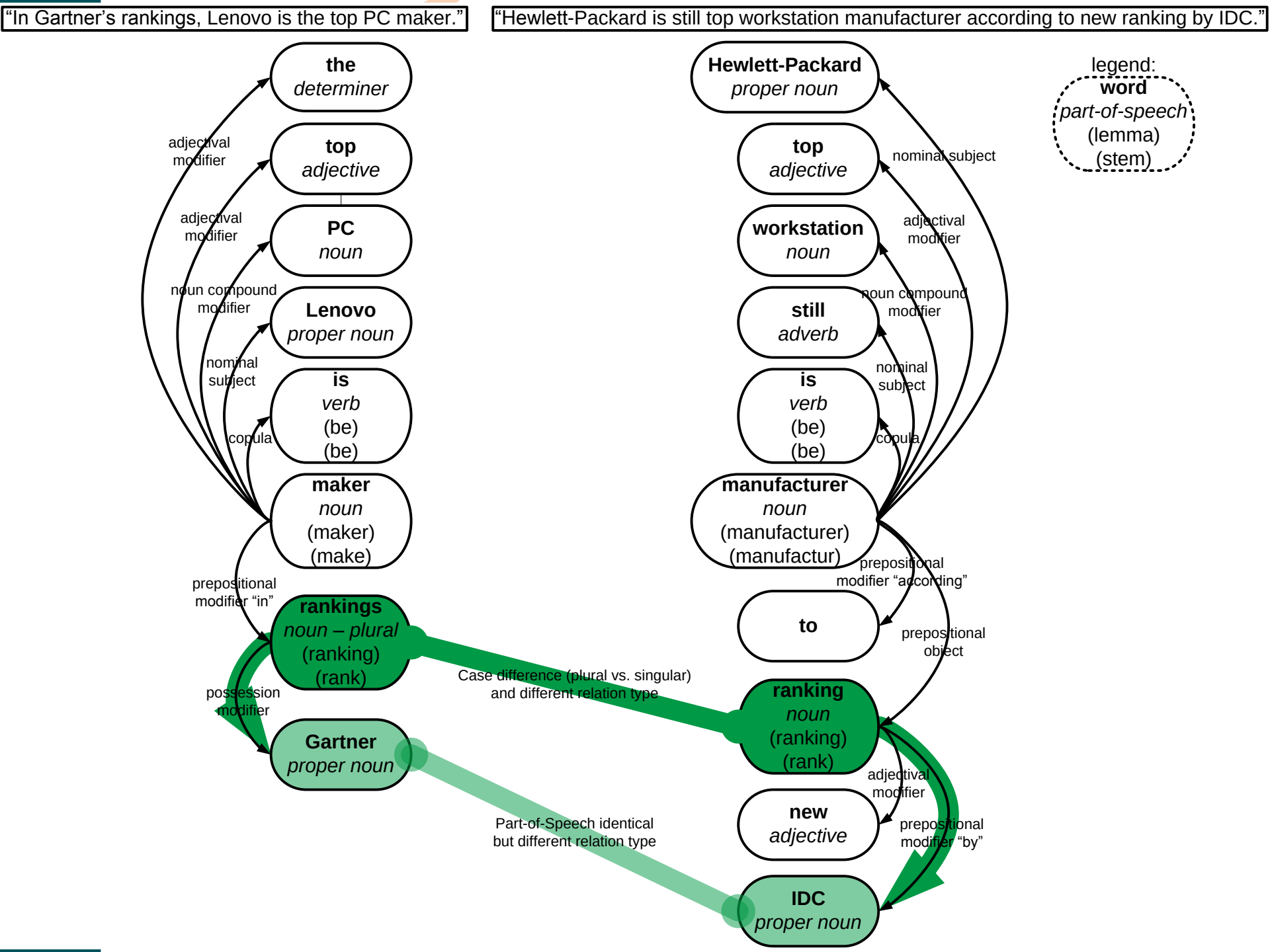
- Nodes are compared on:
  - Basic and full part-of-speech (POS) label
  - Stem, lemma, and fully inflected word
- If POS is the same, but word is not then check for:
  - Synonymy
  - Hypernymy (1 / steps in hypernym tree)
- Correct for word frequency

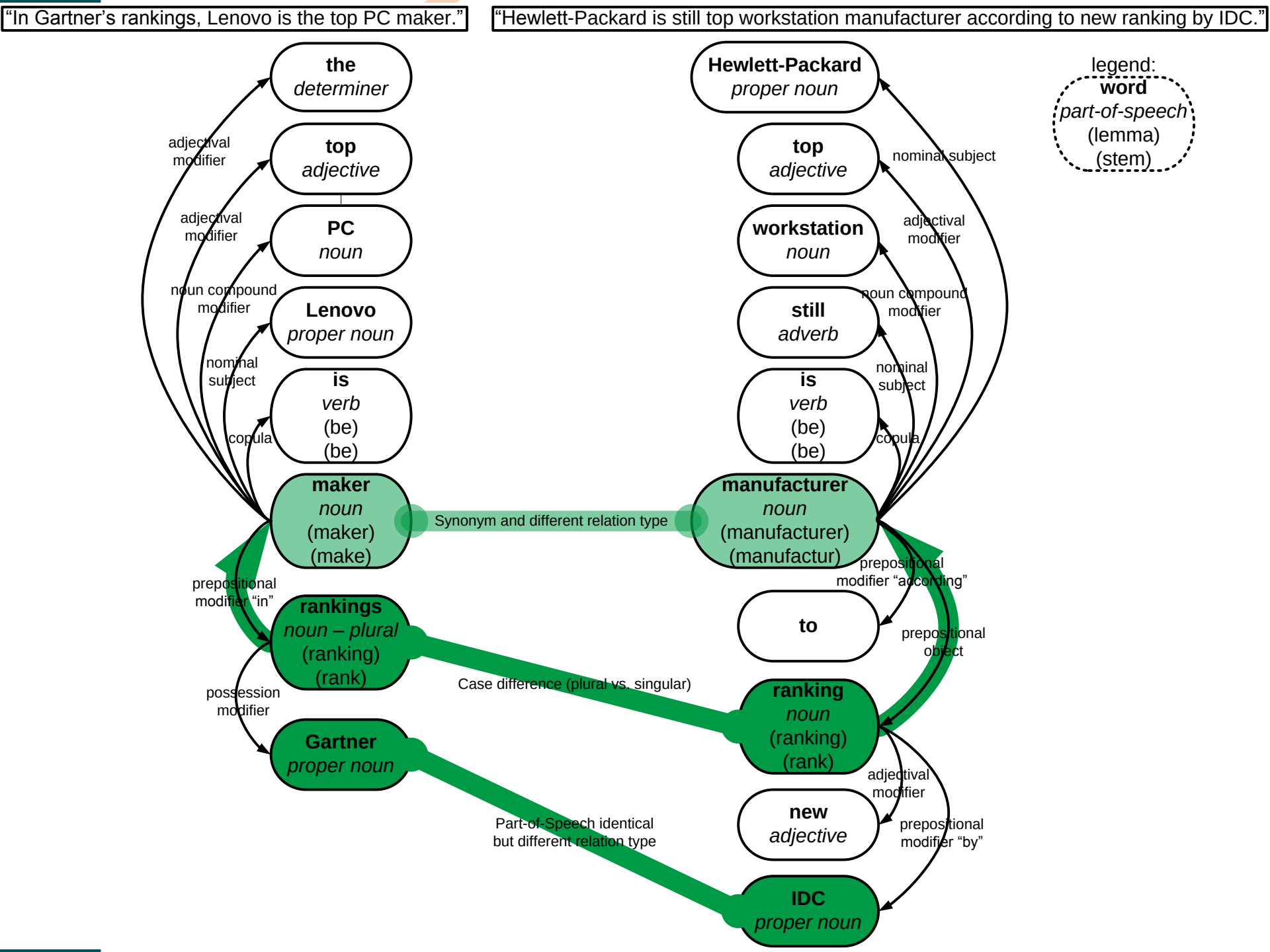
Erasmus Universiteit Rotterdam

Graph comparison

- We can recursively go through both graphs
- Compare nodes and edges to assign score
- However, a starting position within both graphs is needed:
  - Using all possibilities is inefficient
  - Always starting at root is inaccurate
  - Use index of stemmed words (nouns/verbs)
- Only the best scoring starting position is kept

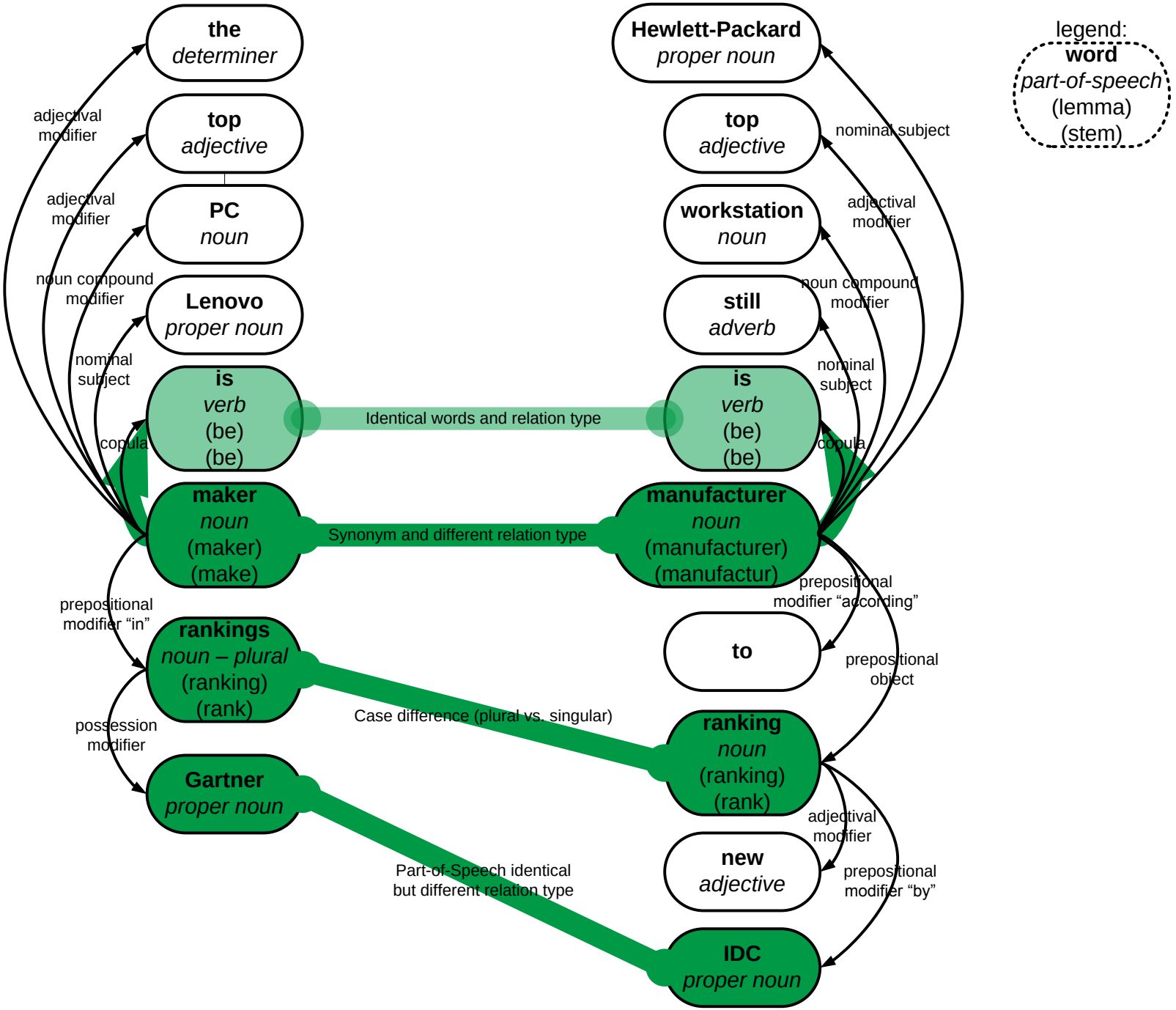

“In Gartner’s rankings, Lenovo is the top PC maker.”
“Hewlett-Packard is still top workstation manufacturer according to new ranking by IDC.”
legend:
word
part-of-speech
(lemma)
(stem)
the
determiner
top
adjective
PC
noun
Lenovo
proper noun
is
verb
(be)
(be)
maker
noun
(maker)
(make)
rankings
noun – plural
(ranking)
(rank)
Gartner
proper noun
adjectival modifier
adjectival modifier
noun compound modifier
nominal subject
copula
prepositional modifier “in”
possession modifier
Hewlett-Packard
proper noun
top
adjective
workstation
noun
still
adverb
is
verb
(be)
(be)
manufacturer
noun
(manufacturer)
(manufactur)
to
ranking
noun
(ranking)
(rank)
new
adjective
IDC
proper noun
nominal subject
adjectival modifier
noun compound modifier
nominal subject
copula
prepositional modifier “according”
prepositional object
adjectival modifier
prepositional modifier “by”
Case difference (plural vs. singular) and different relation type


“In Gartner’s rankings, Lenovo is the top PC maker.”
“Hewlett-Packard is still top workstation manufacturer according to new ranking by IDC.”
legend:
word
part-of-speech
(lemma)
(stem)
the
determiner
top
adjective
PC
noun
Lenovo
proper noun
is
verb
(be)
(be)
maker
noun
(maker)
(make)
rankings
noun – plural
(ranking)
(rank)
Gartner
proper noun
Hewlett-Packard
proper noun
top
adjective
workstation
noun
still
adverb
is
verb
(be)
(be)
manufacturer
noun
(manufacturer)
(manufactur)
to
ranking
noun
(ranking)
(rank)
new
adjective
IDC
proper noun
adjectival modifier
adjectival modifier
noun compound modifier
nominal subject
copula
prepositional modifier “in”
possession modifier
nominal subject
adjectival modifier
noun compound modifier
nominal subject
copula
prepositional modifier “according”
prepositional object
adjectival modifier
prepositional modifier “by”
Case difference (plural vs. singular) and different relation type
Part-of-Speech identical but different relation type


“In Gartner’s rankings, Lenovo is the top PC maker.”
“Hewlett-Packard is still top workstation manufacturer according to new ranking by IDC.”
legend:
word
part-of-speech
(lemma)
(stem)
the
determiner
top
adjective
PC
noun
Lenovo
proper noun
is
verb
(be)
(be)
maker
noun
(maker)
(make)
rankings
noun – plural
(ranking)
(rank)
Gartner
proper noun
Hewlett-Packard
proper noun
top
adjective
workstation
noun
still
adverb
is
verb
(be)
(be)
manufacturer
noun
(manufacturer)
(manufactur)
to
ranking
noun
(ranking)
(rank)
new
adjective
IDC
proper noun
adjectival modifier
adjectival modifier
noun compound modifier
nominal subject
copula
Synonym and different relation type
prepositional modifier “in”
possession modifier
Case difference (plural vs. singular)
Part-of-Speech identical but different relation type
nominal subject
adjectival modifier
noun compound modifier
nominal subject
copula
prepositional modifier “according”
prepositional object
adjectival modifier
prepositional modifier “by”

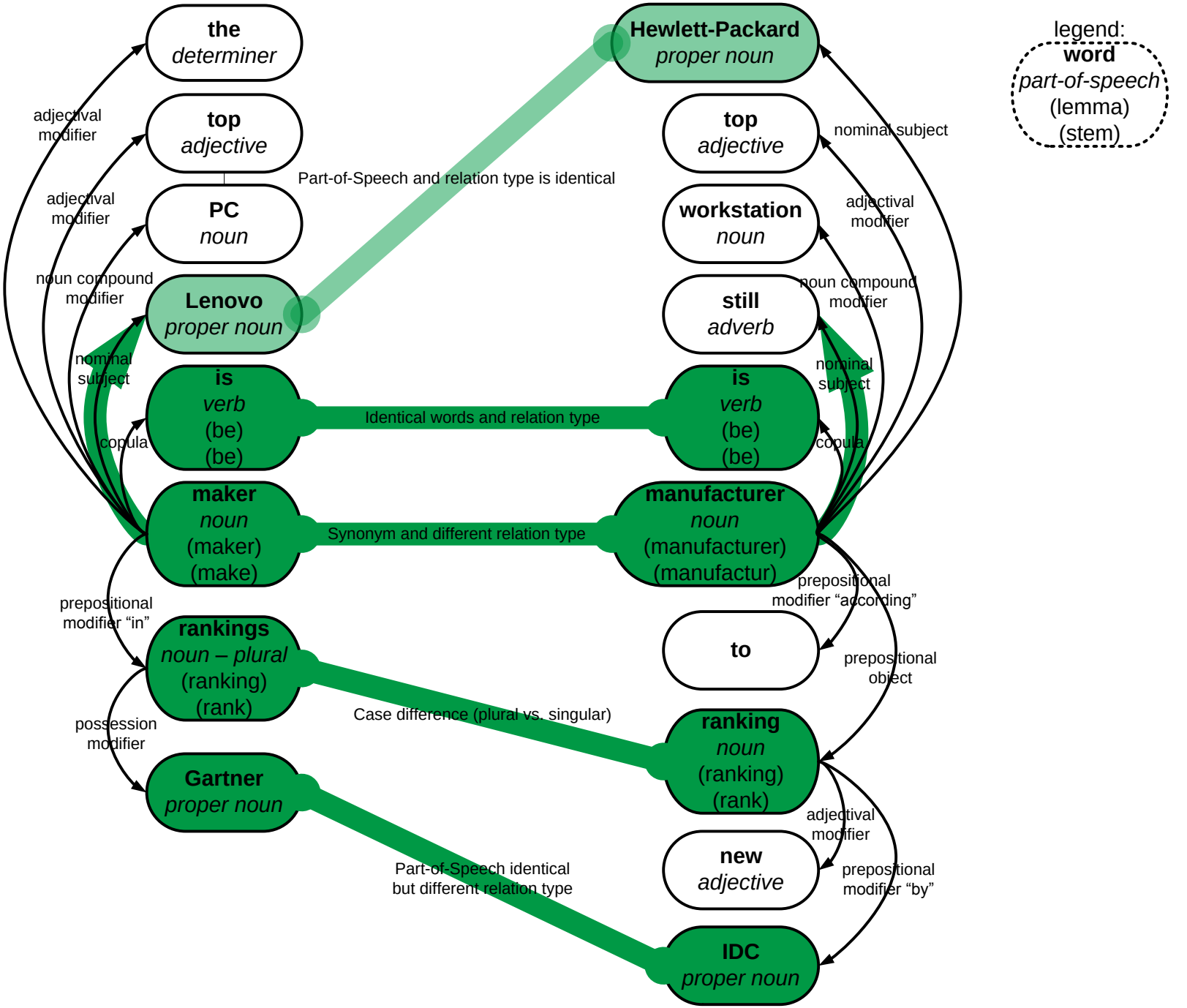
“In Gartner’s rankings, Lenovo is the top PC maker.”

“Hewlett-Packard is still top workstation manufacturer according to new ranking by IDC.”



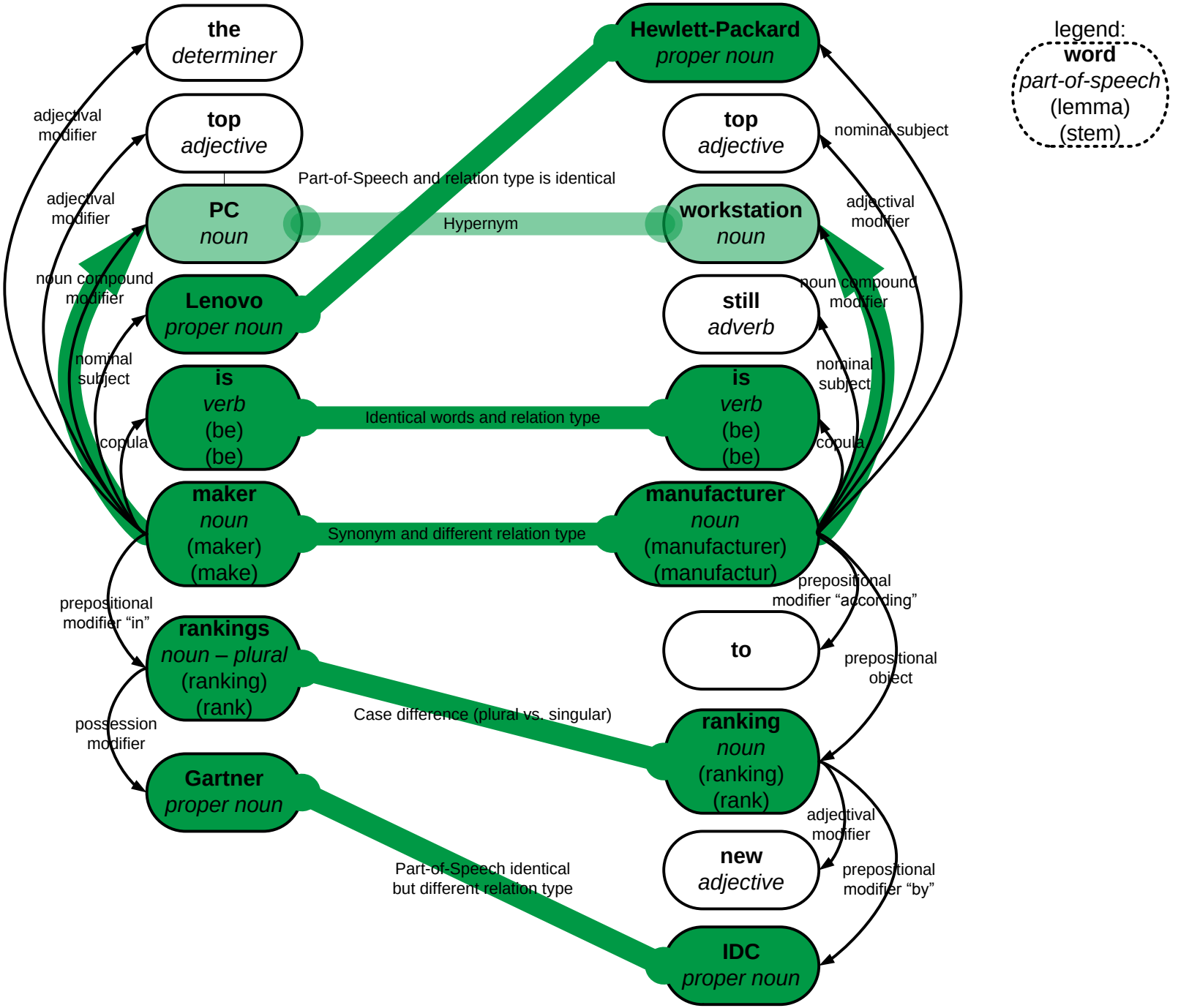
“In Gartner’s rankings, Lenovo is the top PC maker.”

“Hewlett-Packard is still top workstation manufacturer according to new ranking by IDC.”



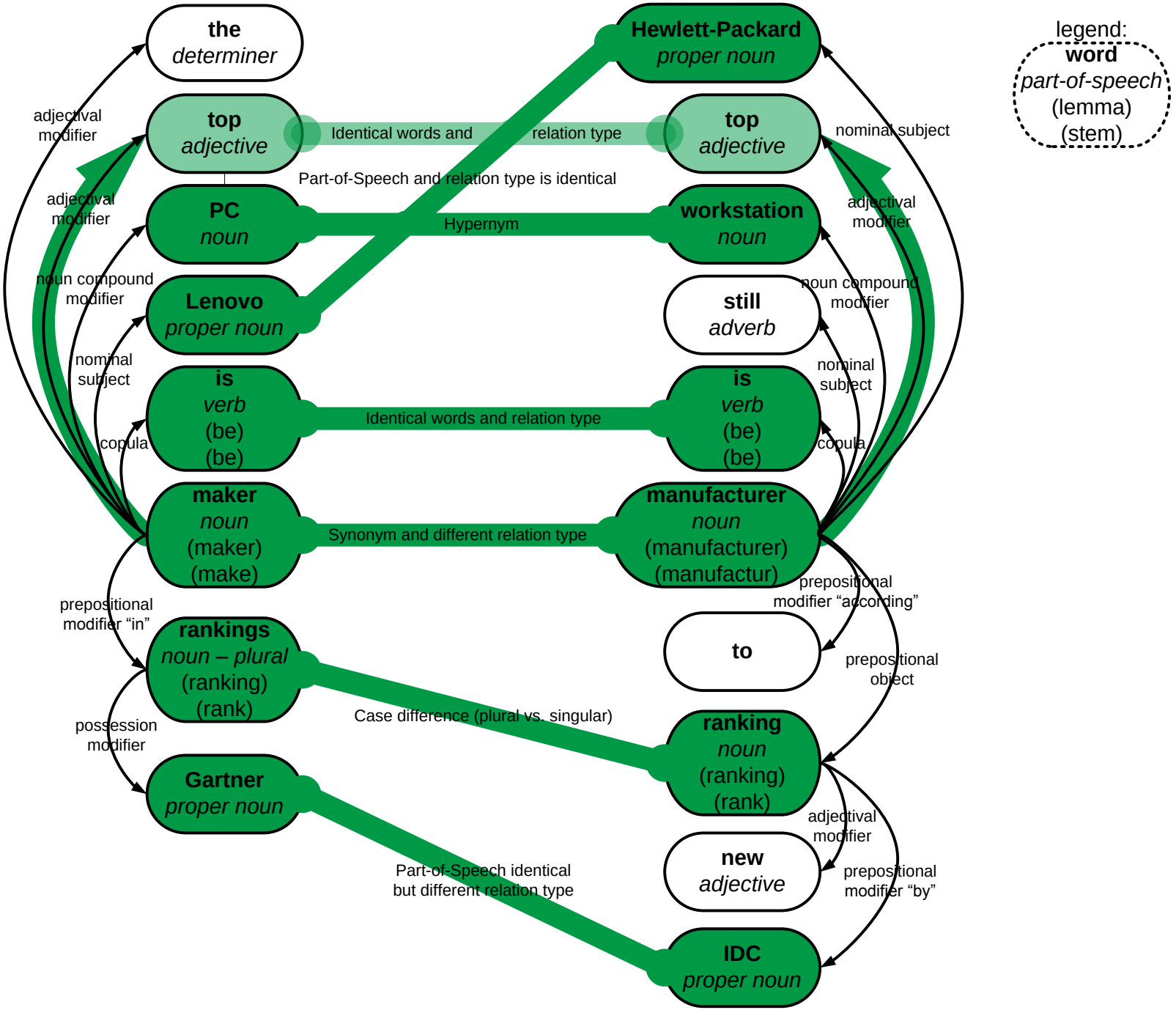
“In Gartner’s rankings, Lenovo is the top PC maker.”

“Hewlett-Packard is still top workstation manufacturer according to new ranking by IDC.”



“In Gartner’s rankings, Lenovo is the top PC maker.”

“Hewlett-Packard is still top workstation manufacturer according to new ranking by IDC.”



Erasmus Universiteit Rotterdam

Data set

- A set of ~1000 sentences
- Extracted from news items
- News items are on roughly the same topic
- 10 sentences are designated as queries
- Three human annotators annotated the similarity between each of the queries and each of the news sentences
- Similarity score of 0,1,2, or 3
- Inter-annotator agreement: 0.1721 std.dev. in score

ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus Universiteit Rotterdam

Score optimization

- Each comparison of two nodes or two edges contributes to total similarity score
- The exact score that each feature can yield is optimized using genetic optimization
- 5 queries and related data are used for training
- Other 5 queries and related data are used for testing
- This is repeated 32 times, with different splits
- For each query a ranked list of sentences is produced according to similarity

ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus Universiteit Rotterdam

Performance

- Results are averages over all 32 splits
- t-statistics are computed over the 32 results for each metric

	TF-IDF mean score	Destiny mean score	rel. improvement	t-test p-value
nDCG	0.238	0.253	11.2%	< 0.001
MAP	0.376	0.424	12.8%	< 0.001
Sp. Rho	0.215	0.282	31.6%	< 0.001

ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus Universiteit Rotterdam

Conclusions

- Our proposed method has several improvements over traditional text searching:
  - By representing text as a graph, the original semantics are preserved, which can be used to leverage search results
  - Words are not only compared lexically, but also semantically, by looking for synonyms and hypernyms

ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus Universiteit Rotterdam

Thank you for your attention!

Questions?

COMMIT/ infiniti

erasmus studio

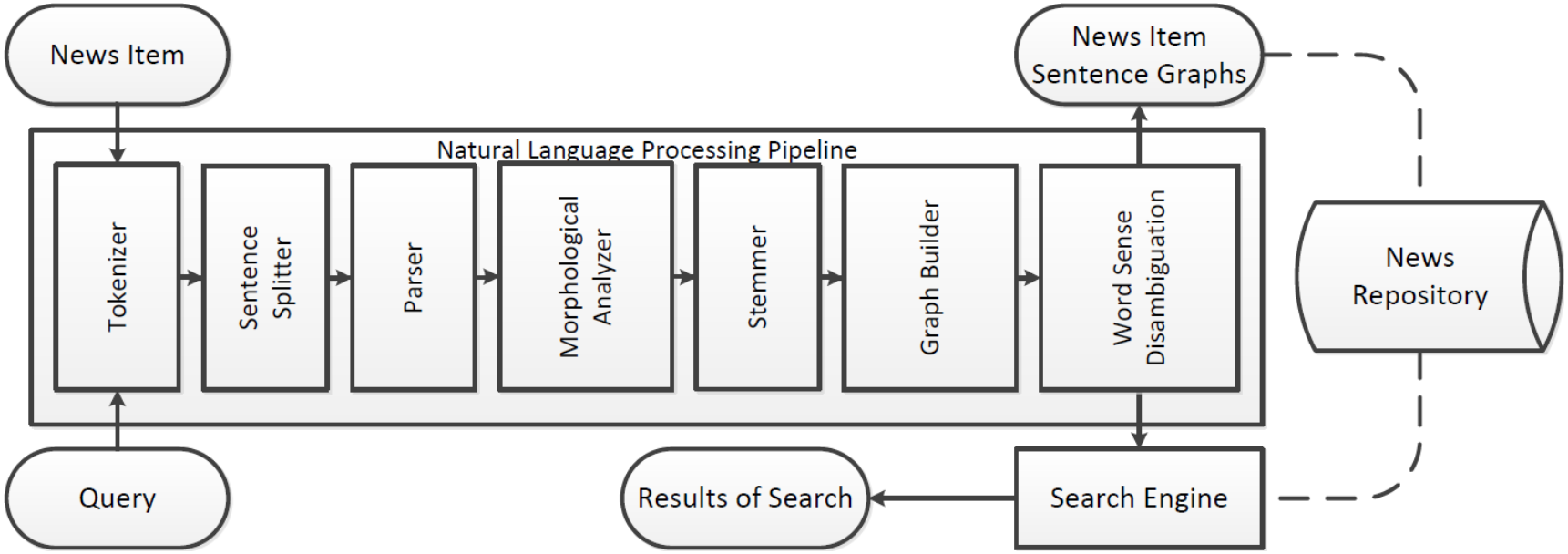
ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus Universiteit Rotterdam

Backup slides

ERASMUS UNIVERSITEIT ROTTERDAM

Pipeline


News Item
News Item Sentence Graphs
Natural Language Processing Pipeline
Tokenizer
Sentence Splitter
Parser
Morphological Analyzer
Stemmer
Graph Builder
Word Sense Disambiguation
News Repository
Query
Results of Search
Search Engine

Erasmus Universiteit Rotterdam

Evaluating ranked lists

- Three metrics: MAP, Spearman's Rho, and nDCG
- MAP measures to what extent the top of the ranking contains only similar/relevant items
  - MAP assumes binary similarity
  - System outputs real-valued similarity scores
  - Converted to binary using cut-off value(s)
  - Cut-off values from 0 to 3 with stepsize 0.1
  - Reported MAP score is average of these

ERASMUS UNIVERSITEIT ROTTERDAM

Evaluating ranked lists

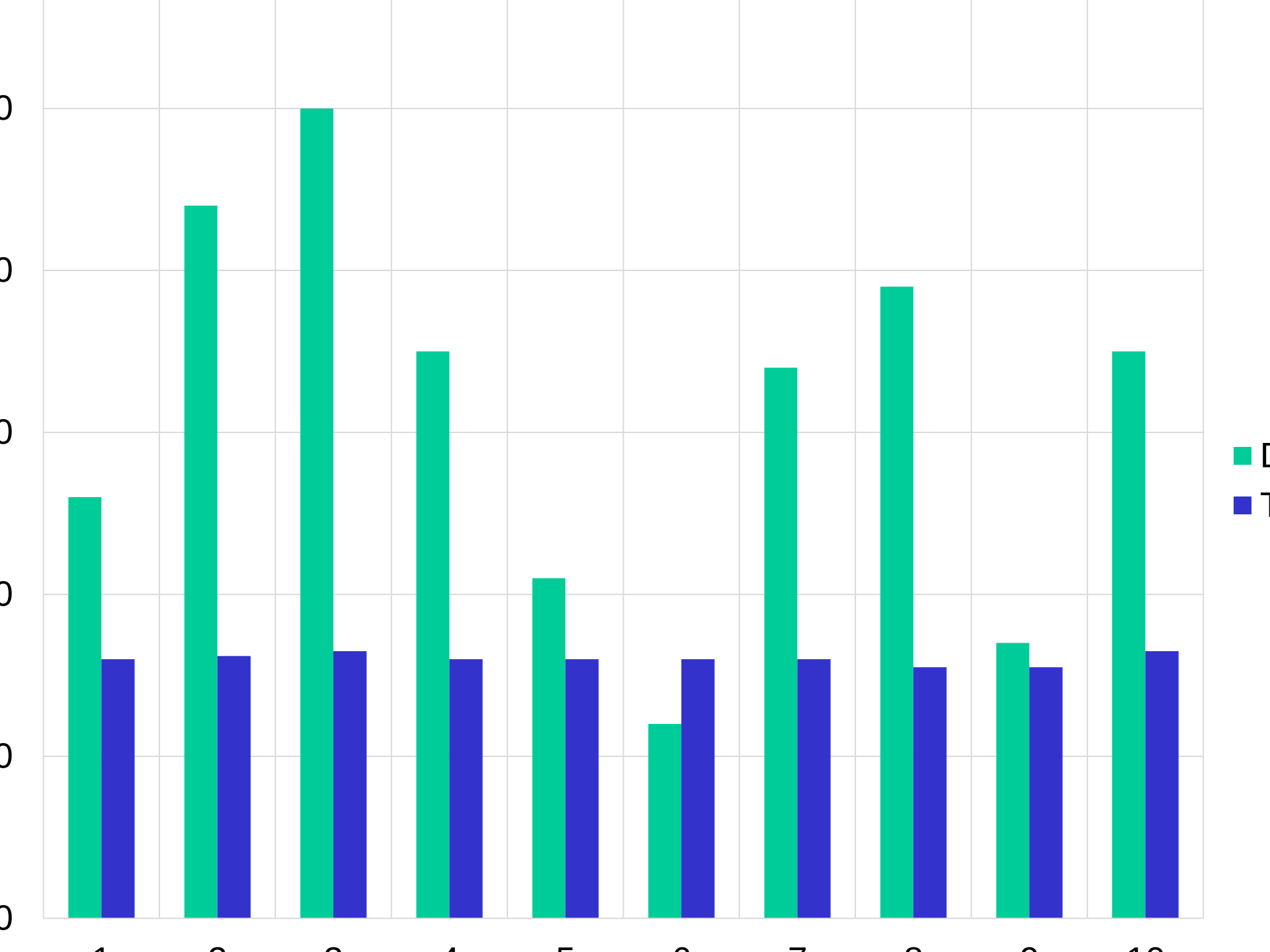
- Spearman's Rho measures correlation of whole list
  - Only top part of results is used in practice
- nDCG measures whether the most similar items are in the top of the ranking
  - Every result contributes its similarity value to final score, discounted by position in ranking
  - Most appropriate
  - Focuses on top part of the ranking
  - Uses real-valued similarity values

Erasmus Universiteit Rotterdam

Scalable?

- Linear in the number of sentences
- Graph comparison is a large ‘constant’ factor
- Depends on:
  - # nodes in query
  - # edges in query
  - Average # nodes in sentences
  - Average # edges in sentences

ERASMUS UNIVERSITEIT ROTTERDAM



Erasmus Universiteit Rotterdam

Open Issues

- More intelligent way to find start positions
- Co-reference resolution
- Non-literal expressions
- Mitigate problems with varying graph sizes
  - “Microsoft is expanding its online corporate offerings to include a full version of Office”
  - “Microsoft includes Office into its online corporate offerings”

ERASMUS UNIVERSITEIT ROTTERDAM