

Paving the Way to a Large-scale Pseudosense-annotated Dataset

Mohammad Taher Pilehvar
Roberto Navigli



SAPIENZA
UNIVERSITÀ DI ROMA

The problem:

Paucity of manually-annotated data

- POS tagged sentences
- Treebanks
- Sense-annotated data
 - SemCor (Miller et al., 1993)
 - MASC (Ide et al., 2010)

A Solution:

Automatic generation of sense-annotated data

- Bootstrapping (Yarowsky, 1995)
- Exploiting parallel data (Chan and Ng, 2005)
- Topic signatures (Martínez et al., 2008)
- Crowdsourcing (Snow et al., 2008)
- Pseudowords (Gale et al., 1992, Schütze, 1992)

What is a pseudoword?

What is a pseudoword?

airplane

river

What is a pseudoword?

airplane

river

airplaneriver

What is a pseudoword?

airplane

river

airplane*river

What is a pseudoword?

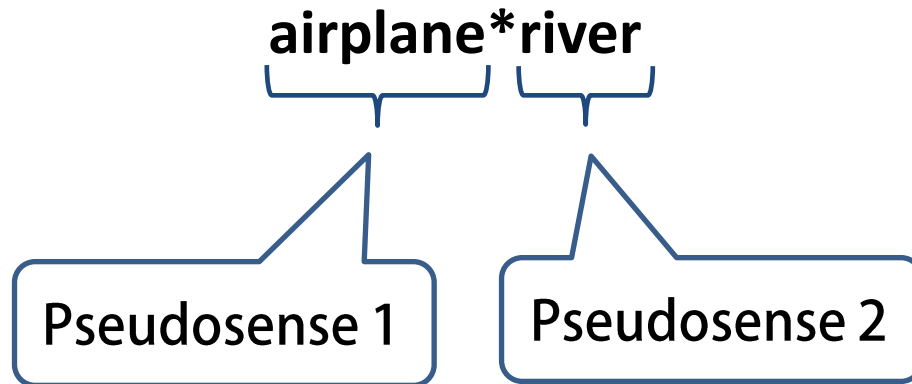
airplane

river

airplane*river

Pseudosense 1

Pseudosense 2



How can pseudowords be used to generate annotated data?

How can pseudowords be used to generate annotated data?

airplane*river

How can pseudowords be used to generate annotated data?

airplane*river

The Wright brothers invented the *airplane*.
The Nile is the longest *river* in the world.

How can pseudowords be used to generate annotated data?

airplane*river

The Wright brothers invented the **airplane**.
The Nile is the longest **river** in the world.

How can pseudowords be used to generate annotated data?

airplane*river

airplane*river



The Wright brothers invented the **airplane**.
The Nile is the longest **river** in the world.

How can pseudowords be used to generate annotated data?

airplane*river

airplane*river



The Wright brothers invented the *airplane*.
The Nile is the longest *river* in the world.



airplane*river

How can pseudowords be used to generate annotated data?

airplane*river

airplane*river



The Wright brothers invented the **airplane**.
The Nile is the longest **river** in the world.



airplane*river



The Wright brothers invented the **airplane*river**.
The Nile is the longest **airplane*river** in the world.

Applications of pseudowords

- Evaluation of
 - Word Sense Disambiguation
Gale et al. (1992), and Schütze (1992)
 - Word Sense Induction
Bordag (2006), and Di Marco and Navigli (2013)
 - Selectional Preferences
Erk (2007), Bergsma et al. (2008), and Chambers and Jurafsky (2010)
 - Information Retrieval
Schütze and Pederson (1995), Sanderson and Rijsbergen (1999)

Some constraints on pseudosenses

- Monosemy

They pulled the canoe up on the **bank***airplane .

Some constraints on pseudosenses

- Monosemy

They pulled the canoe up on the **bank***airplane .

- Sufficient frequency

By 1905, the Wright Flyer III was capable of fully controllable, stable airplane for substantial periods. The Wright brothers credited Otto Lilienthal as a major inspiration for their decision to pursue manned flight.

In 1906, Alberto Santos Dumont made what was claimed to be the first airplane flight unassisted by catapult and set the first world record recognized by the Aéro-Club de France by flying 220 metres (720 ft) in less than 22 seconds. It had movable tail surfaces controlling both yaw and pitch, a form of roll control supplied either by wing warping or by ailerons and controlled by its pilot with a joystick and rudder bar. It was an important predecessor of his later Bleriot XI Channel-crossing aircraft of the summer of 1909.

World War II served as a testbed for the use of the airplane as a weapon. Airplane demonstrated its potential as mobile observation platforms, then proved themselves to be machines of war capable of causing casualties to the enemy. The earliest known aerial victory with a synchronized machine gun-armed fighter aircraft occurred in 1915, by German Luftstreitkräfte Leutnant Kurt Wintgens.

Alcock and Brown crossed the Atlantic non-stop for the first time in 1919. The first international commercial flights took place between the United States and Canada in 1919. Airplane had a presence in all the major battles of World War II. They were an essential component of the military strategies of the period, such as the German Blitzkrieg or the American and Japanese aircraft carrier campaigns of the Pacific War.

Some constraints on pseudosenses

- Monosemy

They pulled the canoe up on the **bank***airplane .

- Sufficient frequency

$$\text{Freq}(\textit{airplane}) = 5$$

By 1905, the Wright Flyer III was capable of fully controllable, stable **airplane** for substantial periods. The Wright brothers credited Otto Lilienthal as a major inspiration for their decision to pursue manned flight.

In 1906, Alberto Santos Dumont made what was claimed to be the first **airplane** flight unassisted by catapult and set the first world record recognized by the Aéro-Club de France by flying 220 metres (720 ft) in less than 22 seconds. It had movable tail surfaces controlling both yaw and pitch, a form of roll control supplied either by wing warping or by ailerons and controlled by its pilot with a joystick and rudder bar. It was an important predecessor of his later Bleriot XI Channel-crossing aircraft of the summer of 1909.

World War II served as a testbed for the use of the **airplane** as a weapon. **Airplane** demonstrated its potential as mobile observation platforms, then proved themselves to be machines of war capable of causing casualties to the enemy. The earliest known aerial victory with a synchronized machine gun-armed fighter aircraft occurred in 1915, by German Luftstreitkräfte Leutnant Kurt Wintgens.

Alcock and Brown crossed the Atlantic non-stop for the first time in 1919. The first international commercial flights took place between the United States and Canada in 1919. **Airplane** had a presence in all the major battles of World War II. They were an essential component of the military strategies of the period, such as the German Blitzkrieg or the American and Japanese aircraft carrier campaigns of the Pacific War.

Why are random pseudowords not good?

airplane*river

- Homonymous distinctions;

cm { Curium
Centimeter

Why are random pseudowords not good?

airplane*river

deficiency

lack, deficiency -- (the state of needing something that is absent or unavailable; "water is the critical deficiency in desert regions")

insufficiency, inadequacy, deficiency -- (lack of an adequate quantity or number; "the inadequacy of unemployment benefits")

We need semantically-aware pseudowords

lack*shortfall

deficiency

lack, deficiency -- (the state of needing something that is absent or unavailable; "water is the critical deficiency in desert regions")

insufficiency, inadequacy, deficiency -- (lack of an adequate quantity or number; "the inadequacy of unemployment benefits")

We need semantically-aware pseudowords

lack*shortfall

- **Category-based pseudowords**

Nakov and Hearst (2003)

- **WordNet-based**

Otrusina and Smrz (2010)

Challenges ahead of pseudoword generation

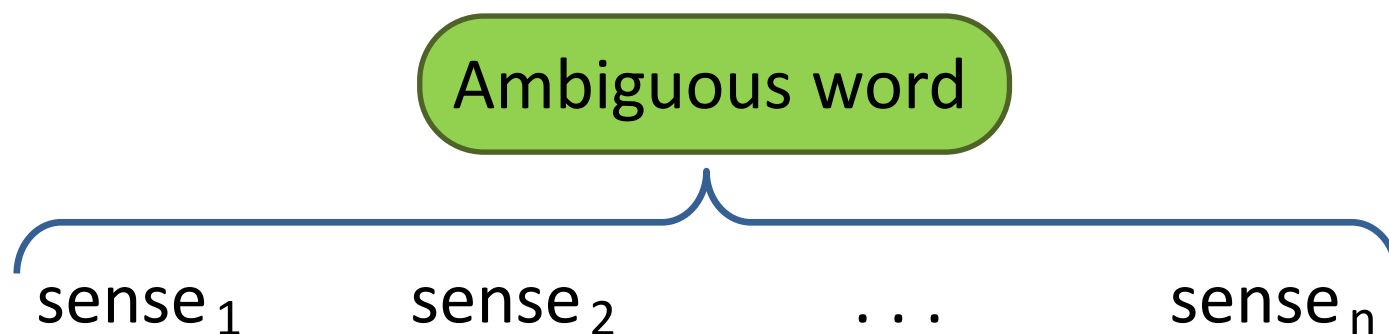
- Semantic awareness
 - E.g.: lack*shortfall

Challenges ahead of pseudoword generation

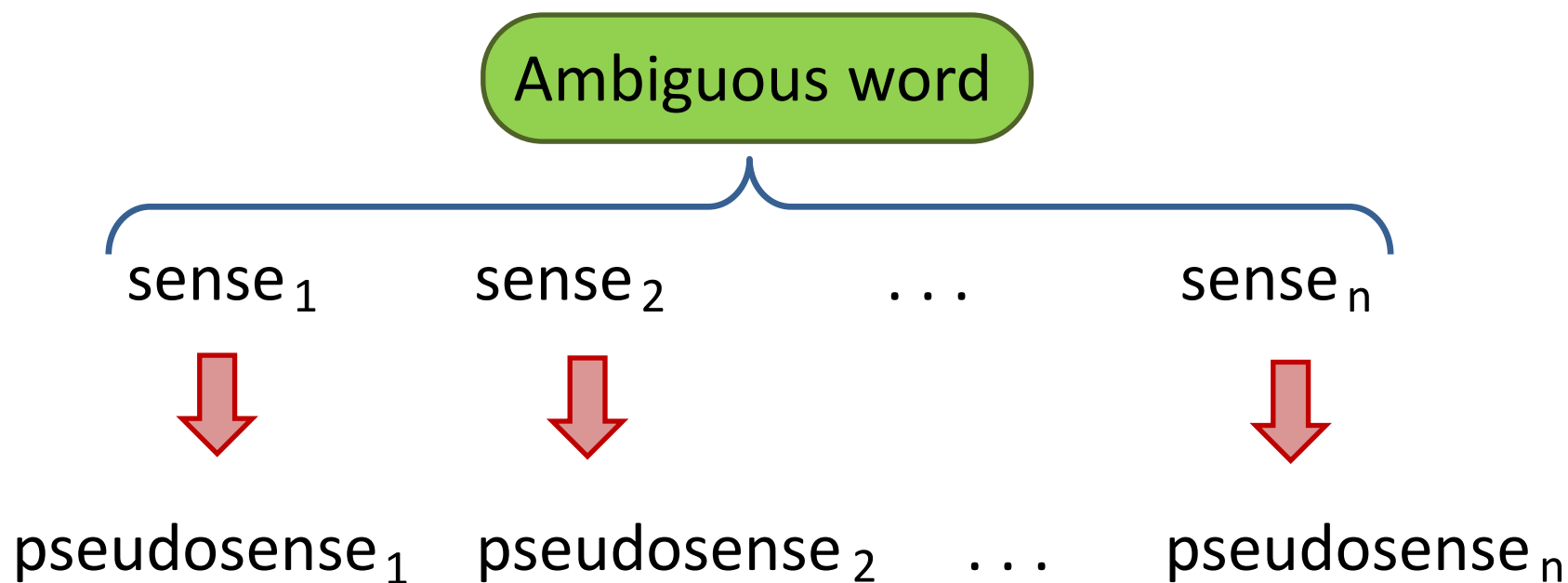
- Semantic awareness
 - E.g.: lack*shortfall
- Coverage
 - Many distinct semantically-aware pseudowords
 - Ideally a pseudowords for each ambiguous word in the lexicon

Our idea: Similarity-based pseudowords

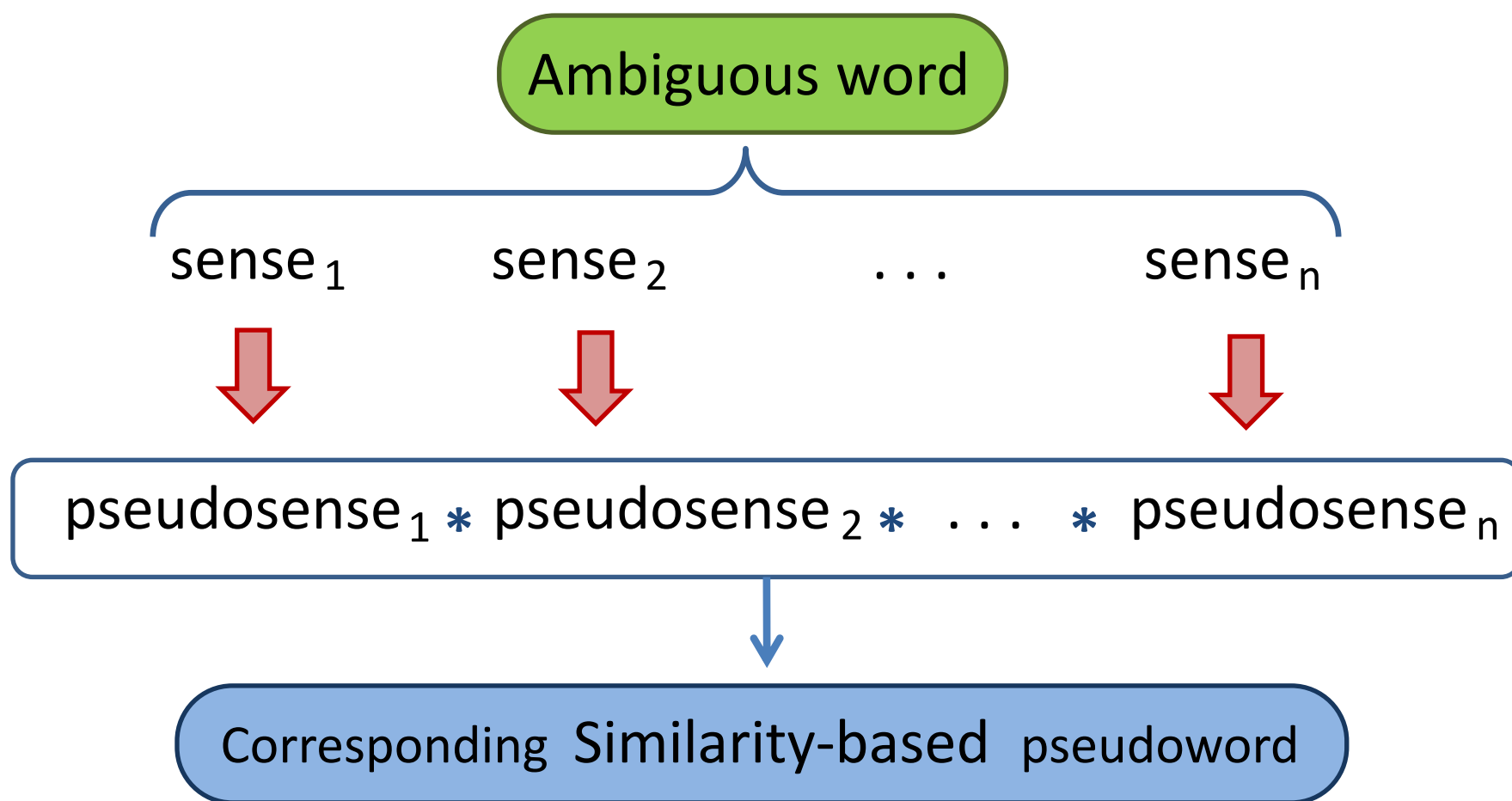
Our idea: Similarity-based pseudowords



Our idea: Similarity-based pseudowords

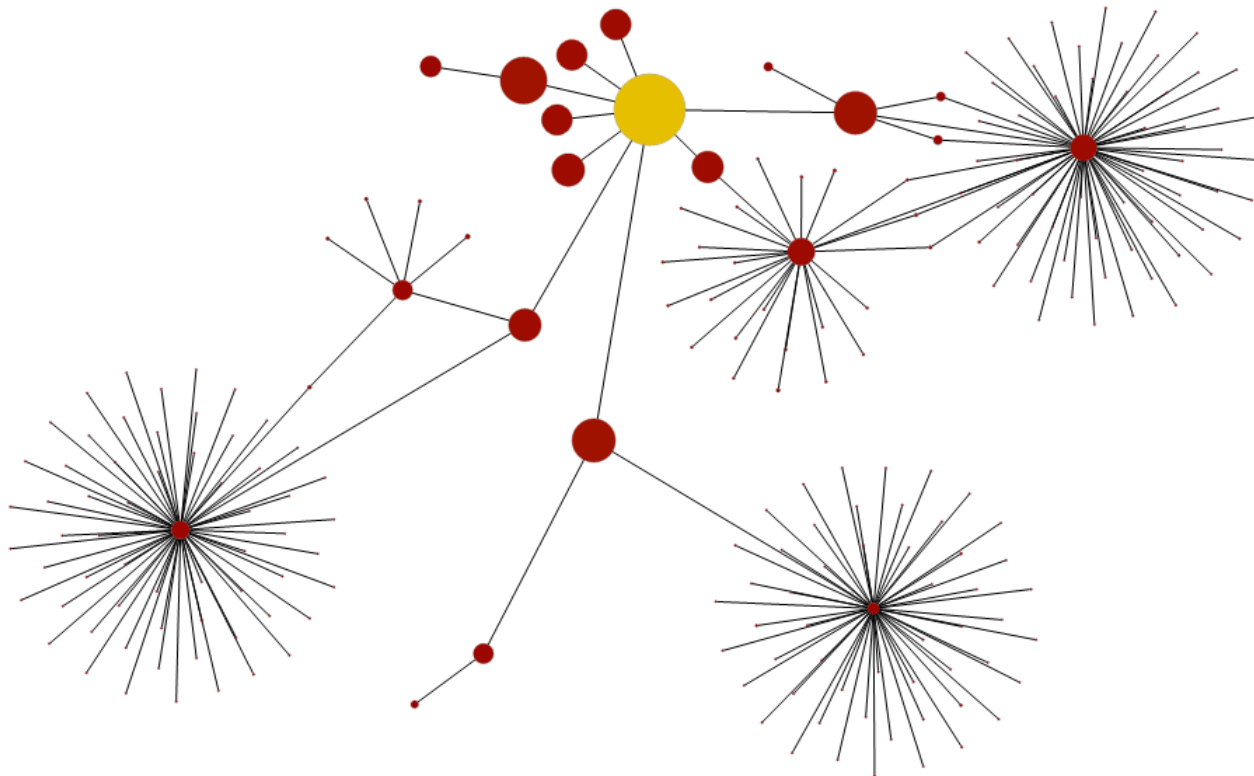


Our idea: Similarity-based pseudowords



Personalized PageRank

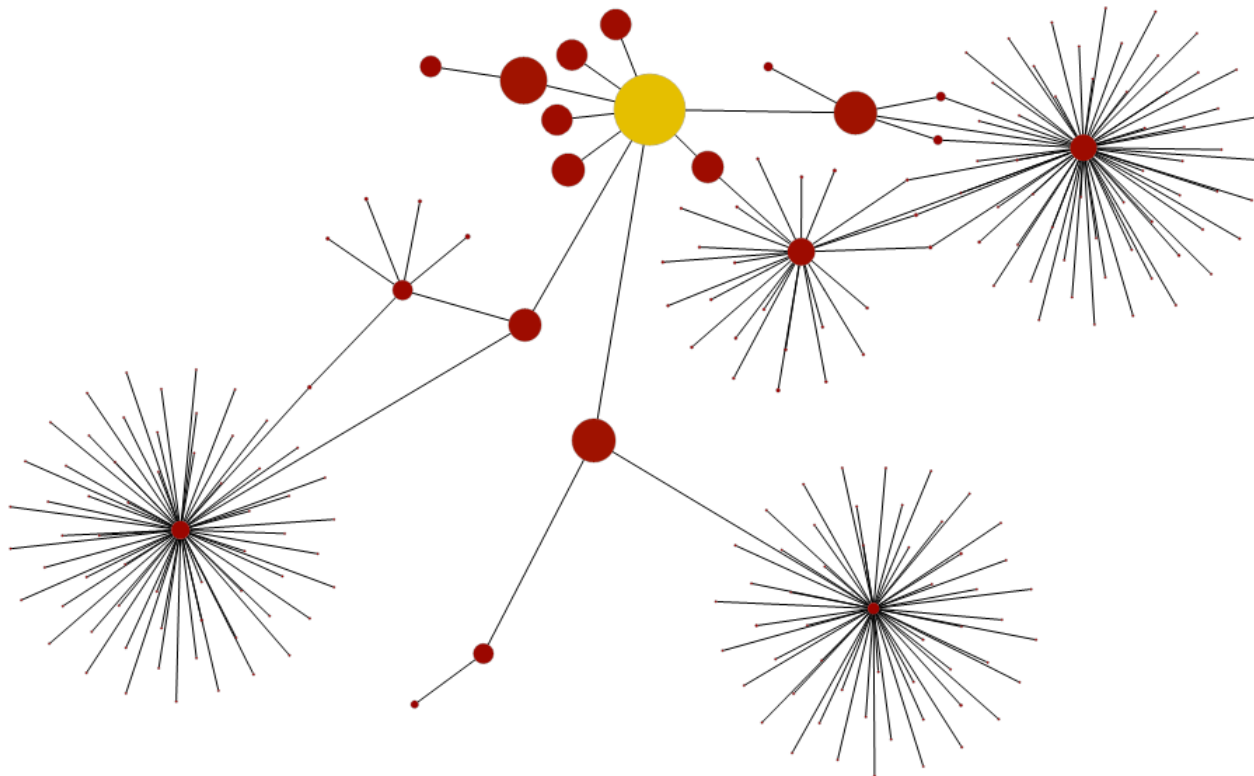
Haveliwala (2002)



Personalized PageRank

Haveliwala (2002)

- Used for semantic similarity by Agirre et al. (2009)

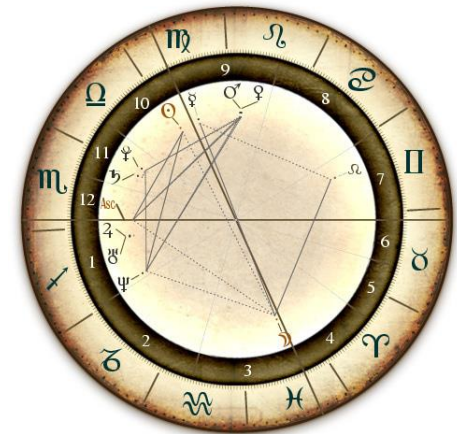


horoscope

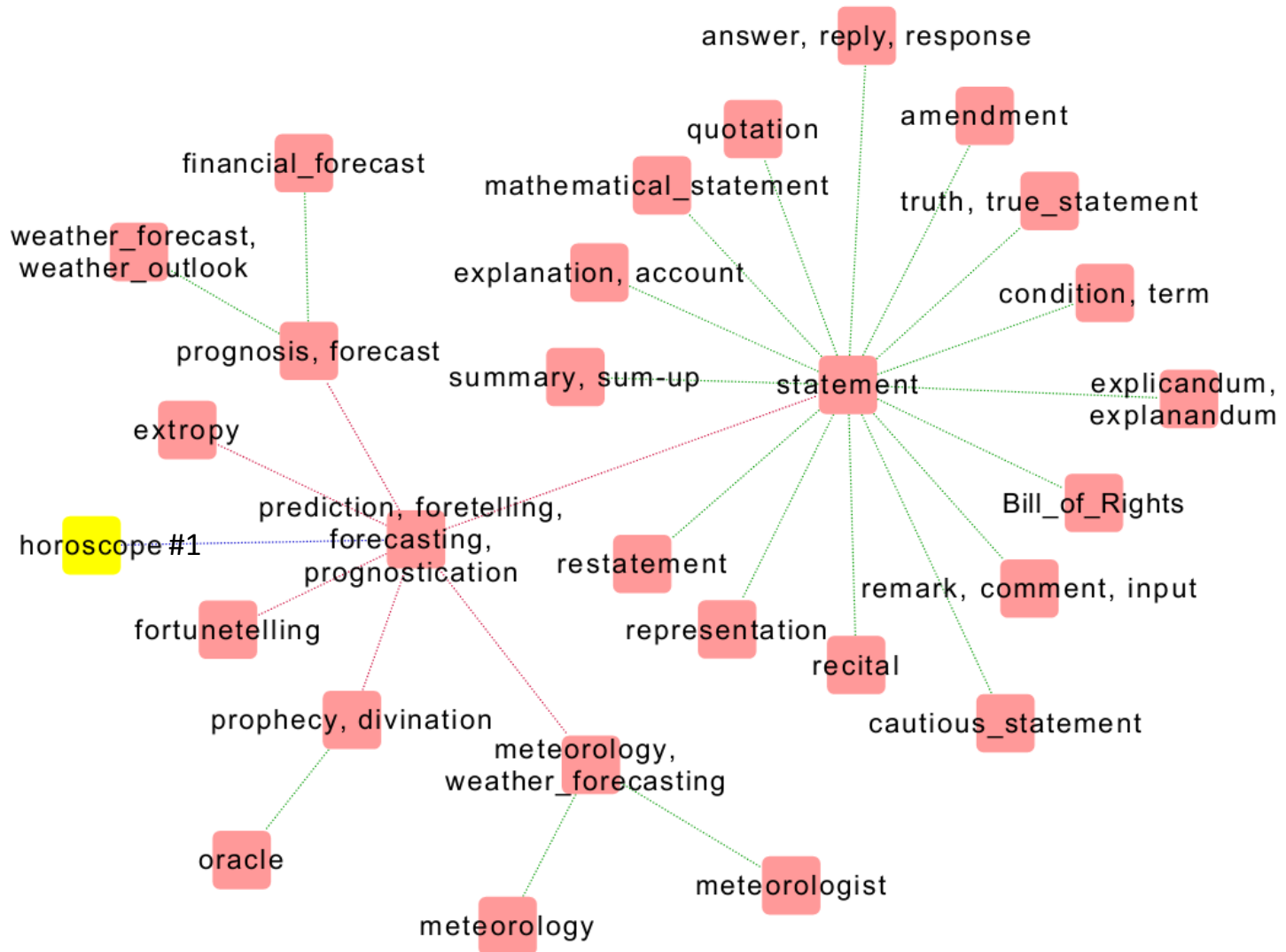
horoscope -- *(a prediction of someone's future based on the relative positions of the planets)*



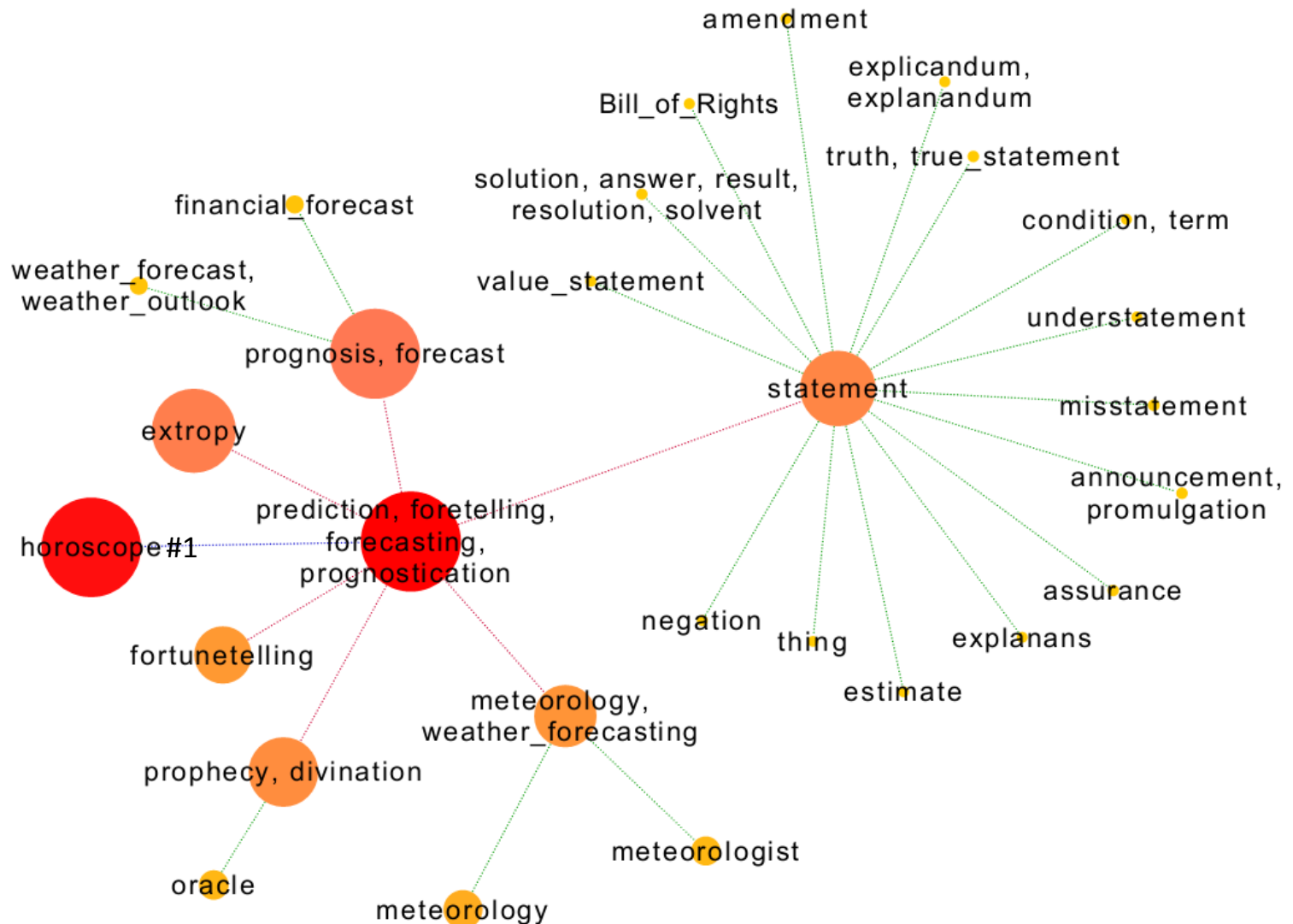
horoscope -- *(a diagram of the positions of the planets and signs of the zodiac at a particular time and place)*



Similarity-based pseudowords



Similarity-based pseudowords



Similarity-based pseudowords

amendment
Bill of Rights explicandum,
explanandum

{prediction, foretelling, forecasting , prognostication}	0.194
{horoscope}	0.174
{prognosis, forecast }	0.031
{ extropy }	0.029
{statement}	0.025
{prophecy, divination}	0.023
{meteorology, weather_forecasting }	0.020
{ fortunetelling }	0.018
{meteorology}	0.011
{oracle}	0.008
.	.
.	.
.	.

prophecy, divination

oracle

meteorology

meteorologist

Similarity-based approach

- ✓ Preserves the **semantic relationship** among senses.

Similarity-based approach

- ✓ Preserves the **semantic relationship** among senses.
- ✓ **Larger search space**, hence higher coverage.

WordNet-based Approach (Otrusina and Smrz, 2010)	vs.	Similarity-based approach
Hyponym Hypernym Meronym Siblings		All WordNet

Similarity-based approach

- ✓ Preserves the **semantic relationship** among senses.
- ✓ **Larger search space**, hence higher coverage.

WordNet-based Approach (Otrusina and Smrz, 2010)	vs.	Similarity-based approach
Hyponym Hypernym Meronym Siblings		All WordNet

- ✓ Does not need **sense-annotated data**.



15,935 pseudowords
for **all** polysemous
nouns in WordNet 3.0



15,935 pseudowords
for **all** polysemous
nouns in WordNet 3.0

Graff and Cieri (2003)



15,935 pseudowords
for **all** polysemous
nouns in WordNet 3.0
(minFreq=**1000**)

Graff and Cieri (2003)



<http://lcl.uniroma1.it/pseudowords/>



15,935 pseudowords
for **all** polysemous
nouns in WordNet 3.0
(minFreq=**1000**)

Graff and Cieri (2003)



<http://lcl.uniroma1.it/pseudowords/>



15,935 pseudowords
for **all** polysemous
nouns in WordNet 3.0
(minFreq=**1000**)

Graff and Cieri (2003)

bernoulli	physicist*mathematician*astronomer
green	greenery*common*labor_leader*green_party*river*golf_course*greens*max
horoscope	forecast*diagram
sunray	sunbeam*vine*sunlight
lifter	athlete*thief



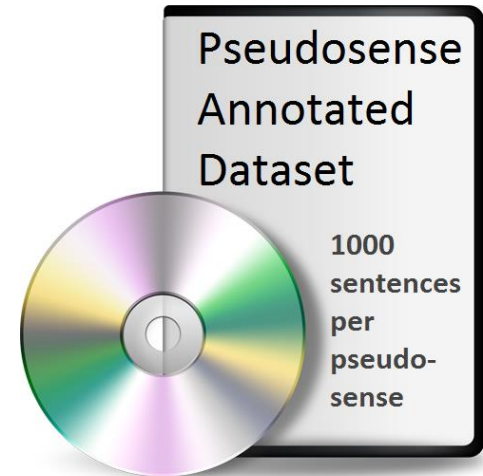
<http://lcl.uniroma1.it/pseudowords/>



15,935 pseudowords
for **all** polysemous
nouns in WordNet 3.0
(minFreq=**1000**)



Graff and Cieri (2003)



bernoulli

physicist*mathematician*astronomer

green

greenery*common*labor_leader*green_party*river*golf_course*greens*max

horoscope

forecast*diagram

sunray

sunbeam*vine*sunlight

lifter

athlete*thief

Evaluating pseudowords

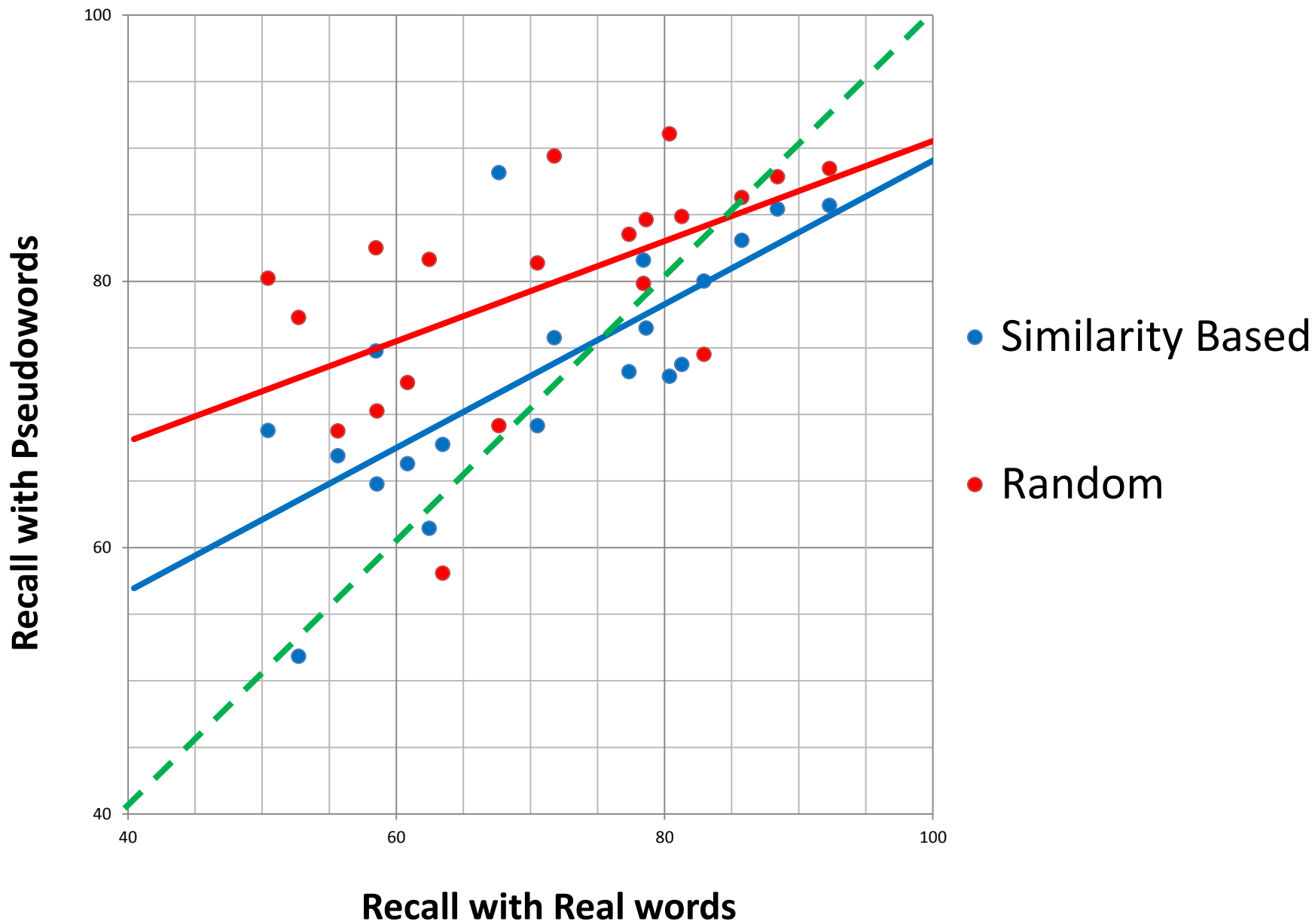
1. Disambiguation difficulty
2. Representativeness of pseudosenses
3. Distinguishability of pseudosenses

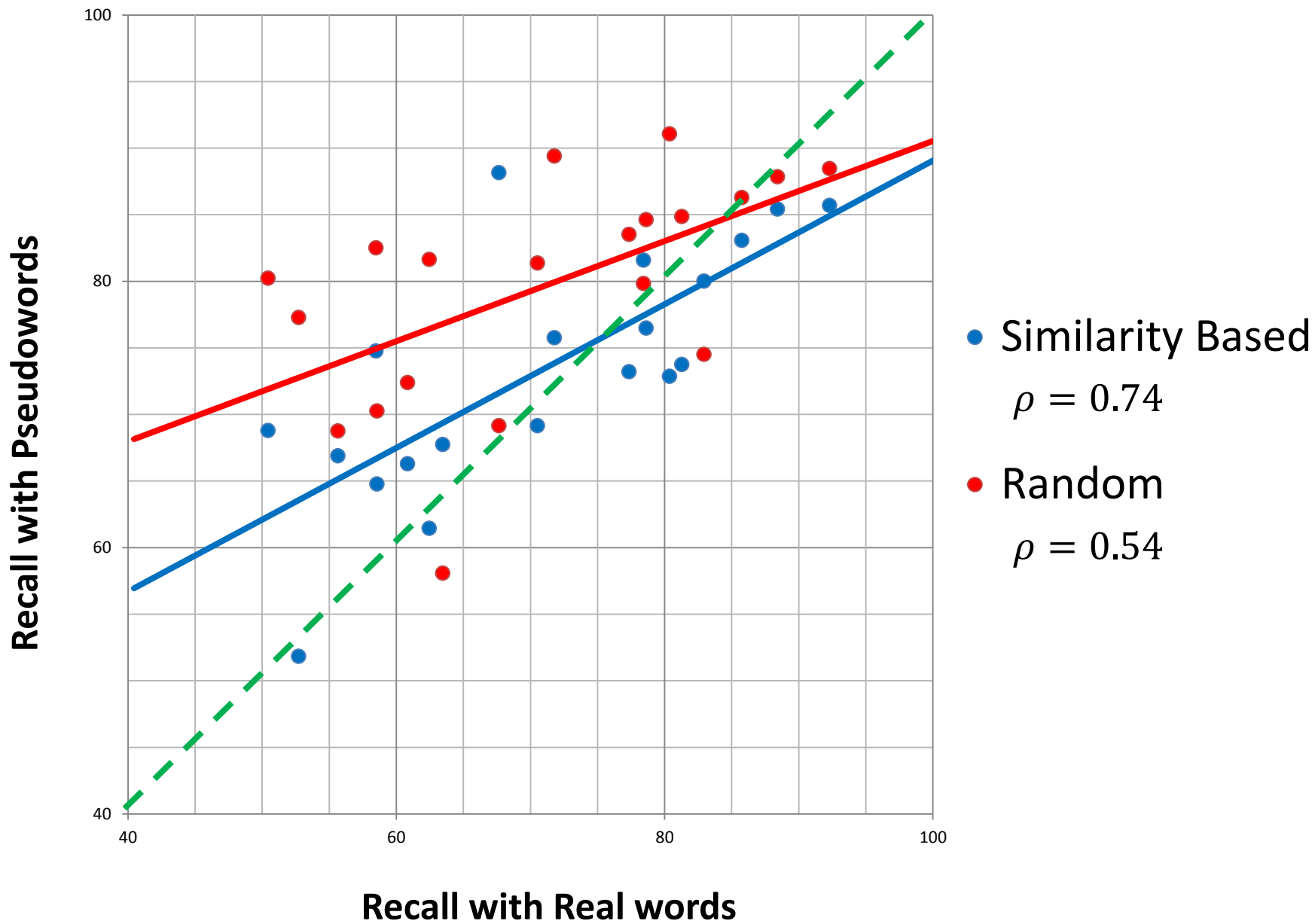
Evaluation I

Disambiguation difficulty of
pseudowords

Disambiguation difficulty of pseudowords

- 20 nouns of the Senseval-3 English Lexical Sample task (Mihalcea et al., 2004)
- Pseudosense-annotated dataset
 - English Gigaword corpus (Graff and Cieri, 2003)
 - Preserved sense distribution
- Baseline: 20 random pseudowords
- WSD System: IMS (ZhiZhong and Ng, 2010)

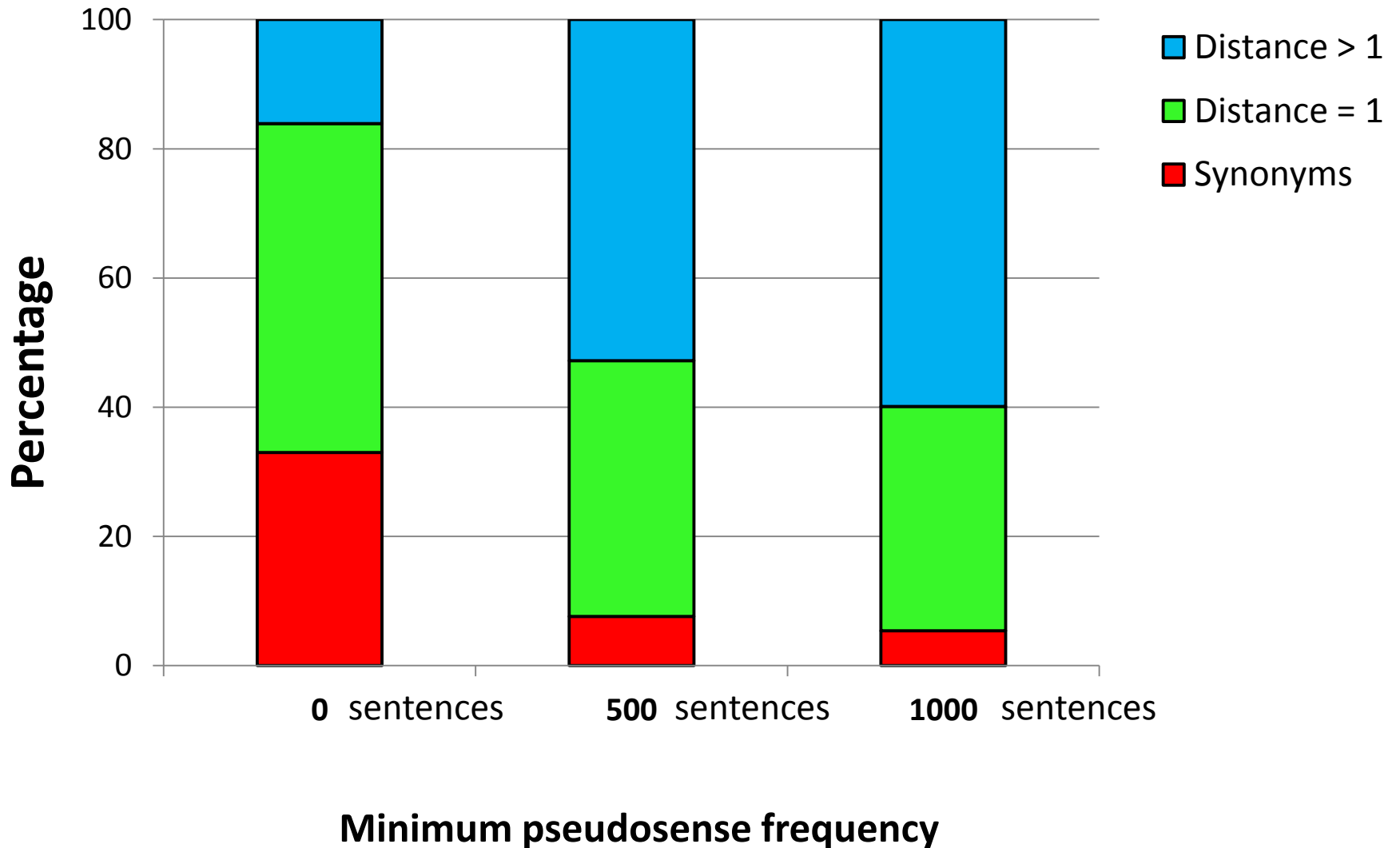




Evaluation 2

Representativeness of pseudosenses

Percentage of similarity-based pseudosenses obtained from different types of WordNet relations



Sampling pseudowords for evaluation

- 110 pseudowords
10 for each polysemy degree 2 to 12
- Only 50 nouns (0.3%) in WordNet 3.0 have polysemy degree > 12

Representativeness of pseudosenses

Representativeness of pseudosenses

representative

Representativeness of pseudosenses

representative

negotiator*spokeperson*congressman*case_in_point

Representativeness of pseudosenses

representative

negotiator*spokeperson*congressman*case_in_point

representative#1



negotiator

representative#2



spokeperson

representative#3



congressman

representative#4



case_in_point

Representativeness of pseudosenses

A person who represents others
representative

negotiator

An advocate who represents someone else's policy
spokeperson, interpreter, representative, voice

spokeperson

A member of the U.S. House of Representatives
congressman, congresswoman, representative

congressman

An item of information that is typical of a group
example, illustration, instance, representative

case_in_point

Representativeness of pseudosenses

A person who represents others
representative

negotiator

An advocate who represents someone else's policy
spokeperson, interpreter, representative, voice

spokeperson

A member of the U.S. House of Representatives
congressman, congresswoman, representative

congressman

An item of information that is typical of a group
example, illustration, instance, representative

case_in_point

1: completely unrelated

2: somewhat related

3: good substitute

4: perfect substitute

Representativeness of pseudosenses

A person who represents others <i>representative</i>	<i>negotiator</i>	3	3
An advocate who represents someone else's policy <i>spokeperson, interpreter, representative, voice</i>	<i>spokeperson</i>	4	4
A member of the U.S. House of Representatives <i>congressman, congresswoman, representative</i>	<i>congressman</i>	4	4
An item of information that is typical of a group <i>example, illustration, instance, representative</i>	<i>case_in_point</i>	4	3
		3.75	3.5

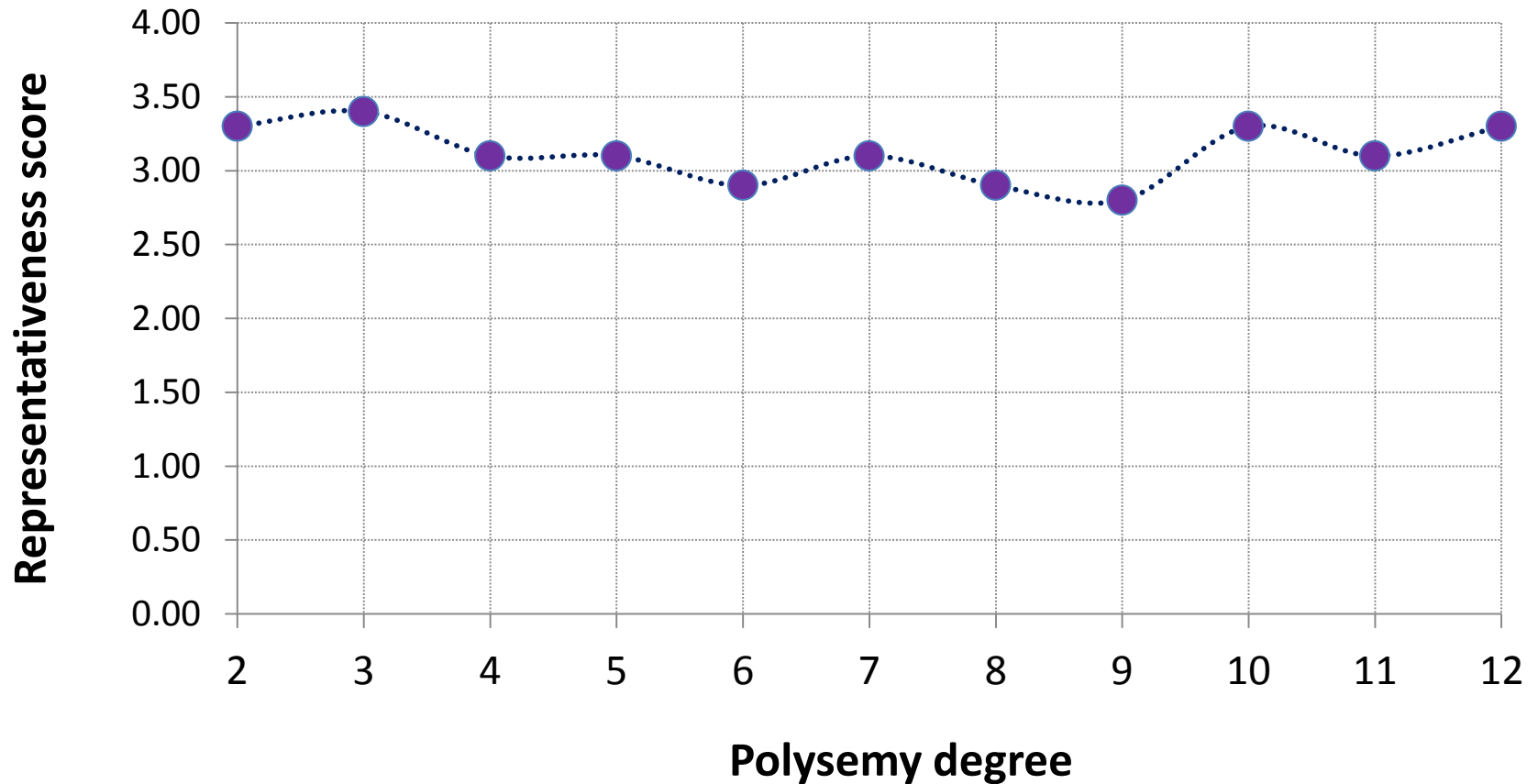
1: completely unrelated

2: somewhat related

3: good substitute

4: perfect substitute

Representativeness of pseudosenses



Evaluation 3

Distinguishability of pseudosenses

Distinguishability of pseudosenses

donor

Distinguishability of pseudosenses

donor

1. donor, giver, presenter,
bestower, conferrer
(person who makes a gift of
property)

2. donor
((medicine) someone who
gives blood or tissue or an
organ to be used in another
person (the host))

Distinguishability of pseudosenses

donor

1. donor, giver, presenter,
bestower, conferrer
(person who makes a gift of
property)

2. donor
((medicine) someone who
gives blood or tissue or an
organ to be used in another
person (the host))

philanthropist*benefactor

Distinguishability of pseudosenses

donor

1. donor, giver, presenter,
bestower, conferrer
(person who makes a gift of
property)

2. donor
((medicine) someone who
gives blood or tissue or an
organ to be used in another
person (the host))

philanthropist*benefactor



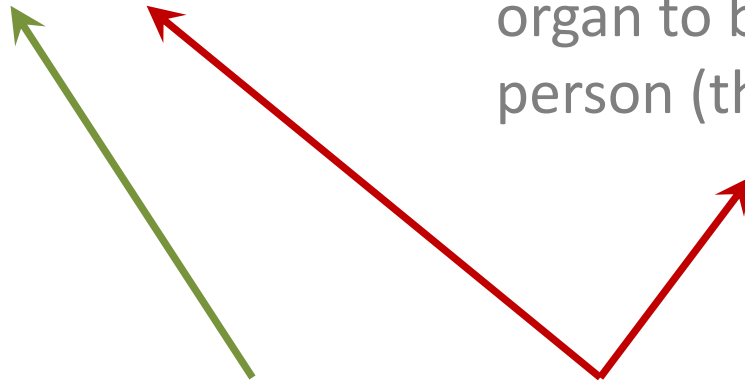
Distinguishability of pseudosenses

donor

1. donor, giver, presenter,
bestower, conferrer
(person who makes a gift of
property)

2. donor
((medicine) someone who
gives blood or tissue or an
organ to be used in another
person (the host))

philanthropist*benefactor



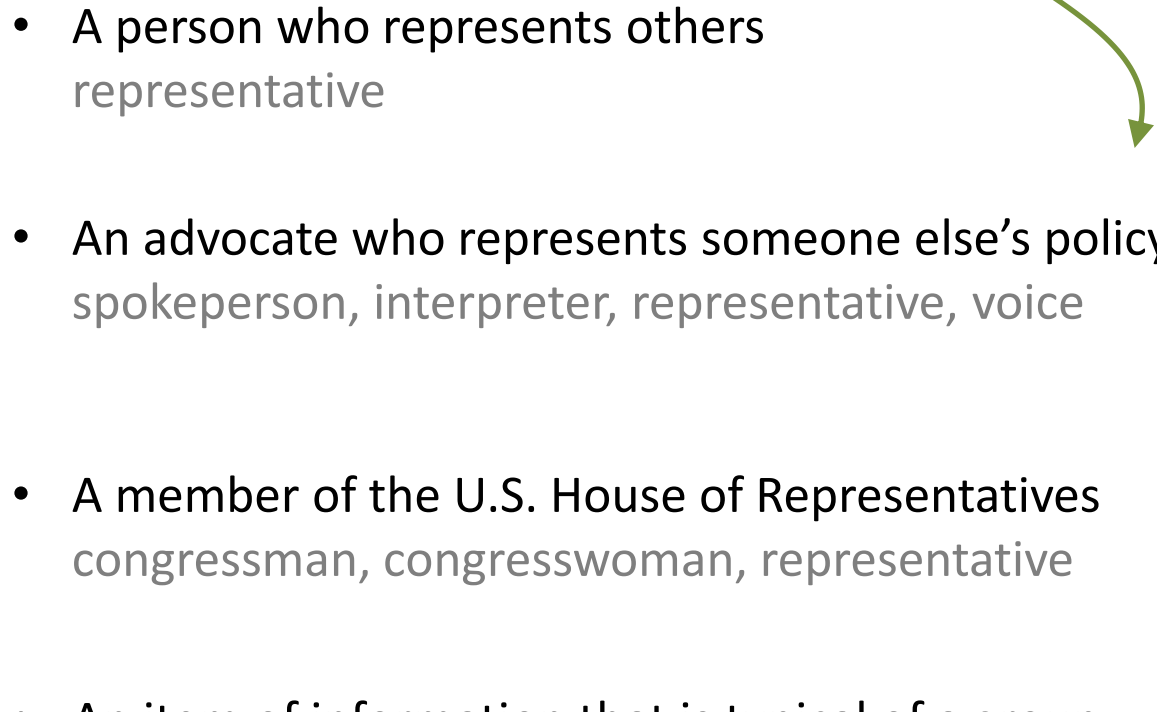
[spokeperson, case_in_point, negotiator, congressman]

- A person who represents others
representative
- An advocate who represents someone else's policy
spokeperson, interpreter, representative, voice
- A member of the U.S. House of Representatives
congressman, congresswoman, representative
- An item of information that is typical of a group
example, illustration, instance, representative

[spokeperson, case_in_point, negotiator, congressman]

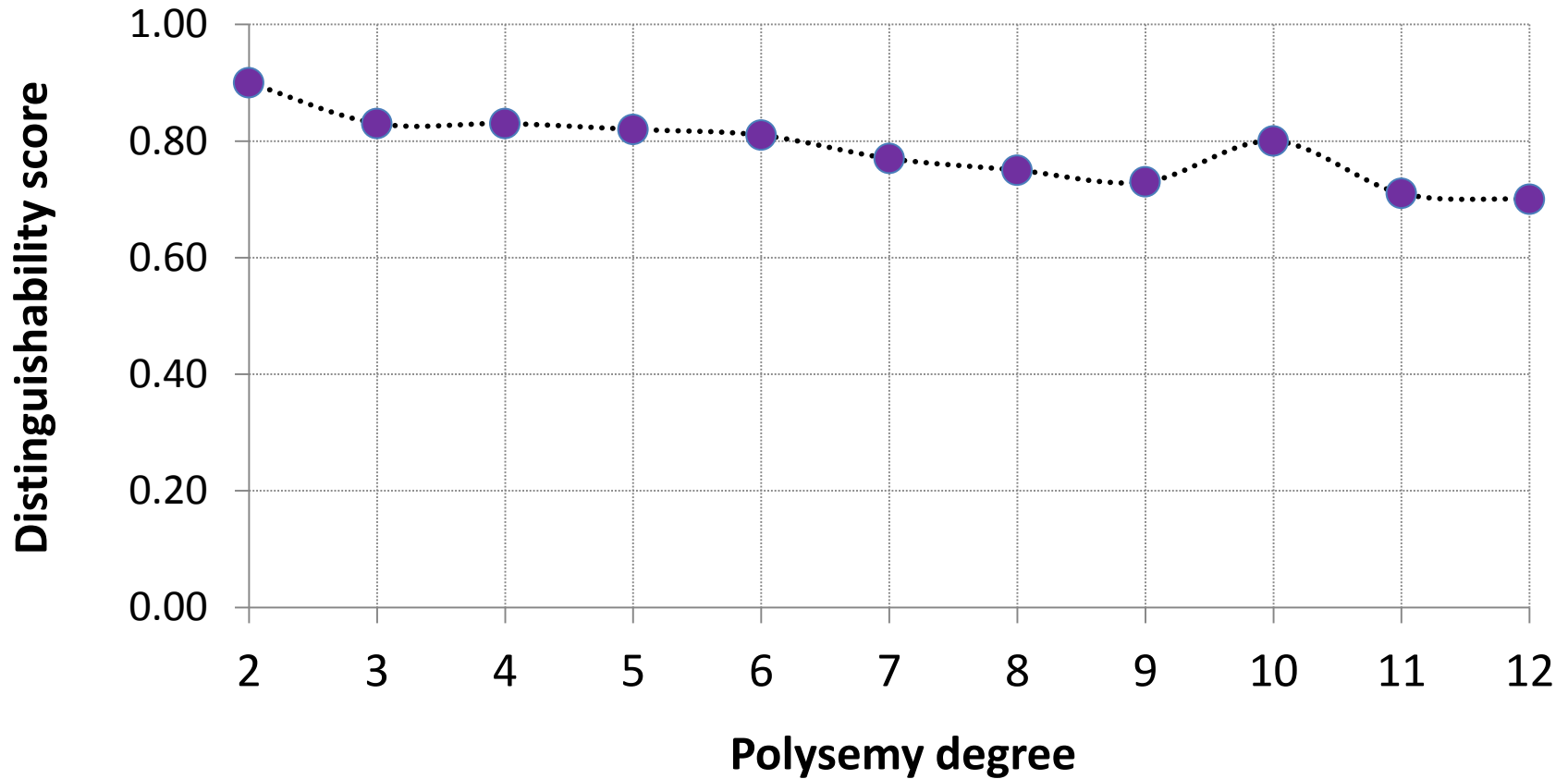
- A person who represents others
representative
- An advocate who represents someone else's policy
spokeperson, interpreter, representative, voice
- A member of the U.S. House of Representatives
congressman, congresswoman, representative
- An item of information that is typical of a group
example, illustration, instance, representative

[spokeperson, case_in_point, negotiator, congressman]

- 
- A person who represents others
representative
 - An advocate who represents someone else's policy
spokeperson, interpreter, representative, voice
 - A member of the U.S. House of Representatives
congressman, congresswoman, representative
 - An item of information that is typical of a group
example, illustration, instance, representative

$4/4 = 1$

Distinguishability scores



Conclusions

- Similarity-based pseudowords
 - Semantic-awareness
 - Coverage
- Three evaluation experiments

Thanks!



<http://lcl.uniroma1.it/pseudowords/>

Category-based Pseudowords

(Nakov and Hearst, 2003)

- MeSH
- Eye

- A01: Body Region

- A09: Sense Organ



thumb

pupils