3rd CAMELEON Workshop

Porto Alegre, 20 December 2012

Distributional Similarity Models

Discovering related terms

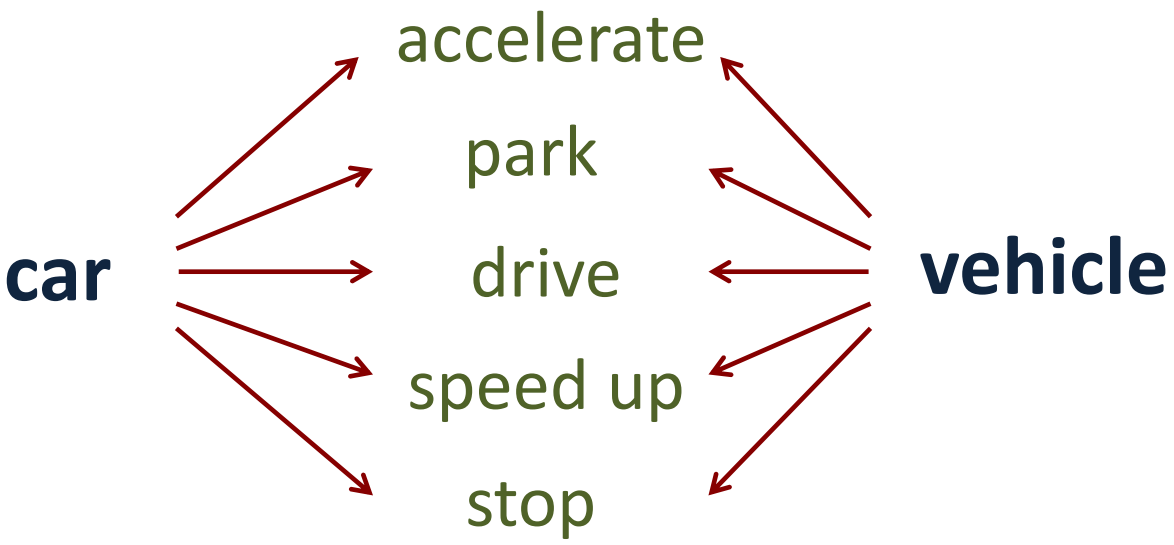
Roger Leitzke Granada

roger.granada@acad.pucrs.br

Pontifical Catholic University of Rio Grande do Sul
Doctoral Program in Computer Science

What are DSMs?

- **Distributional Similarity Models** (DSMs) identify similar words using the **distributional hypothesis** - Similar words appear in similar contexts (Harris, 1954).



Example

- Related terms to ***tezgüino*** (Lin, 1998)*:*

Example

- Related terms to ***tezgüino*** (Lin, 1998)*:*

A bottle of tezgüino is on the table.
Everyone likes tezgüino.
Tezgüino makes you drunk.
We make tezgüino out of corn.

Example

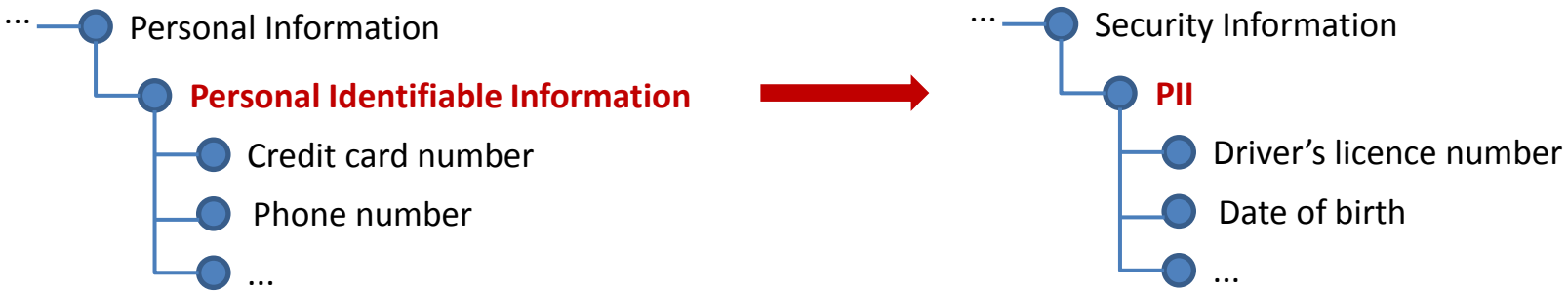
- Related terms to ***tezgüino*** (Lin, 1998)*:*

  A bottle of tezgüino is on the table.
  Everyone likes tezgüino.
  Tezgüino makes you drunk.
  We make tezgüino out of corn.

- **Tezgüino**: beer, wine, vodka etc.

Why should I look for similar terms?

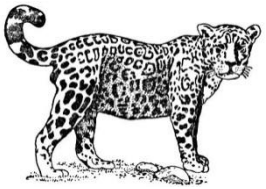
- Align ontology concepts



Why should I look for similar terms?

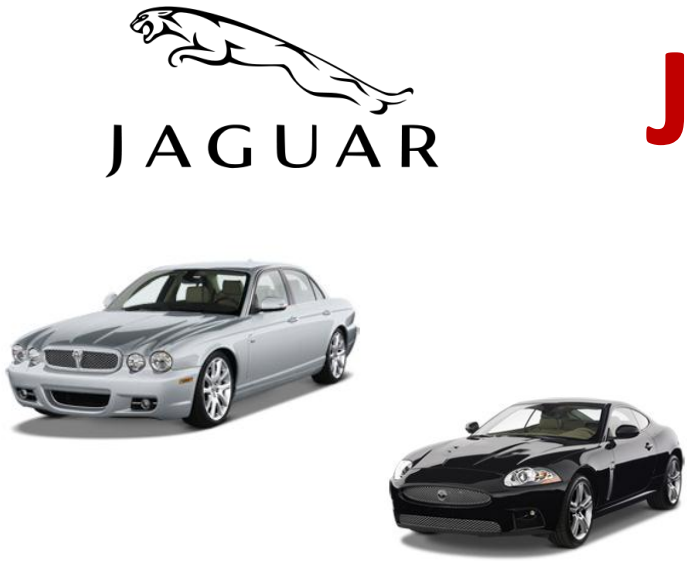
- Word Sense Disambiguation

Jaguar



Why should I look for similar terms?

- Word Sense Disambiguation



Jaguar

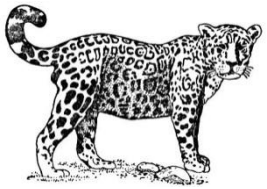
Related terms:
- Car
- Vehicle
- Taxi
- Passenger

Why should I look for similar terms?

- Word Sense Disambiguation

Related terms:
- Animal
- Cheetah
- Panther
- Tiger

Jaguar

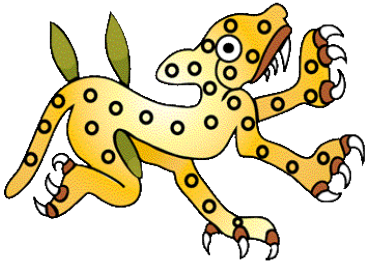


Why should I look for similar terms?

- Word Sense Disambiguation

Related terms :
- Mesoamerican
- Leopard
- Panther
- Maya

Jaguar



How can I find related terms?

- Statistical methods

"You shall know a word by the company it keeps."

John Rupert Firth (1957)

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

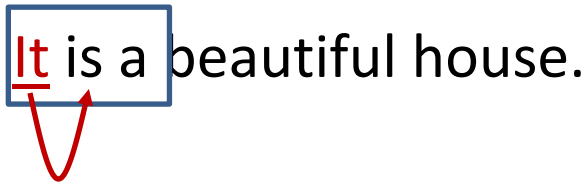
It is a beautiful house.

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words



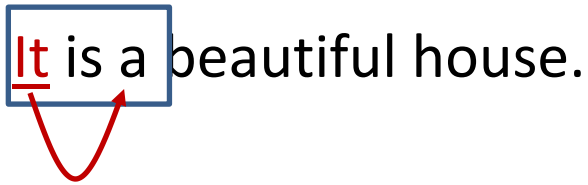
It – is = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words



It – is = 1
It – a = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words

It is a beautiful house.

It – is = 1
It – a = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words

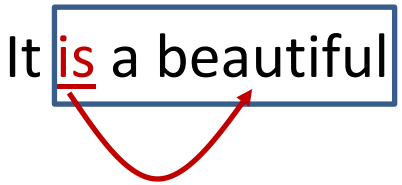
It – is = 1
It – a = 1
Is – a = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words



It – is = 1
It – a = 1
Is – a = 1

Is – beautiful = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words

It is a beautiful house.

It – is = 1
It – a = 1
Is – a = 1

Is – beautiful = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words

It – is = 1
It – a = 1
Is – a = 1

Is – beautiful = 1
A – beautiful = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words

It – is = 1
It – a = 1
Is – a = 1

Is – beautiful = 1
A – beautiful = 1
A – house = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words

It is a beautiful house.

It – is = 1
It – a = 1
Is – a = 1

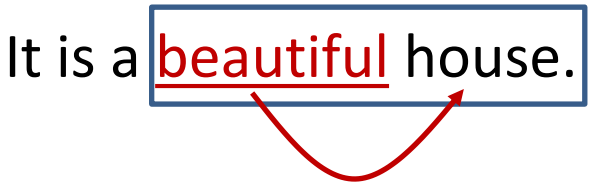
Is – beautiful = 1
A – beautiful = 1
A – house = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window

Example:

- Window = 3 words



It – is = 1
It – a = 1
Is – a = 1

Is – beautiful = 1
A – beautiful = 1
A – house = 1

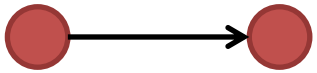
Beautiful – house = 1

How can I find related terms?

- Statistical methods
  - First oder co-occurrences
  - Extract terms in a window
  - Apply the Mutual Information (Church e Hanks, 1990)

$$MI(t_i, t_j) = log_2 \left(\frac{P(t_i, t_j)}{P(t_i)P(t_j)} \right)$$

Example



beautiful house

How can I find related terms?

- Using linguistic features

"Words that occur in the same contexts tend to have similar meanings."

Zellig Harris (1954)

How can I find related terms?

- Using linguistic features
  - Second order co-occurrences

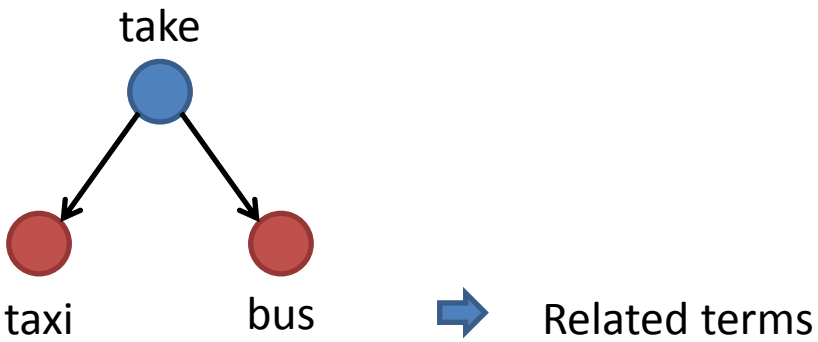
"Words that share **syntactic** contexts tend to have similar meanings."

Gregory Grefenstette (1994)

What is a context?

"Words that share **syntactic** contexts tend to have similar meanings."

Example:
- John took the taxi.
- Mary took the bus to go home.

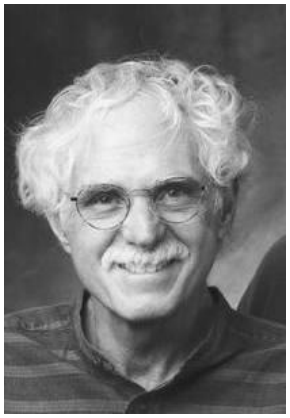


How can I find related terms?

- Using latent relations

"We assume that there is some underlying or 'latent' structure in the pattern of word usage that is partially obscured by the variability of word choice."

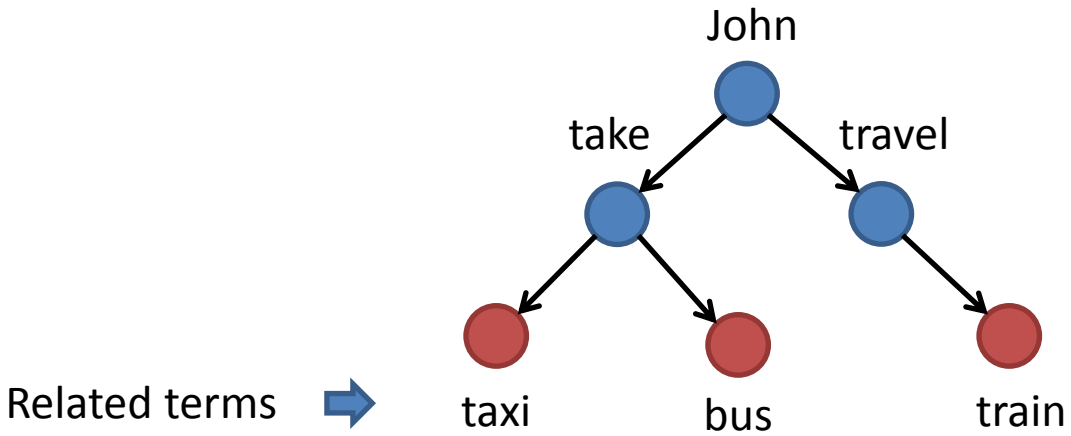
Thomas Landauer and
Susan Dumais (1997)



How can I find related terms?

- Using Latent Semantic Analysis (LSA)
  - Third (or more) order co-occurrence

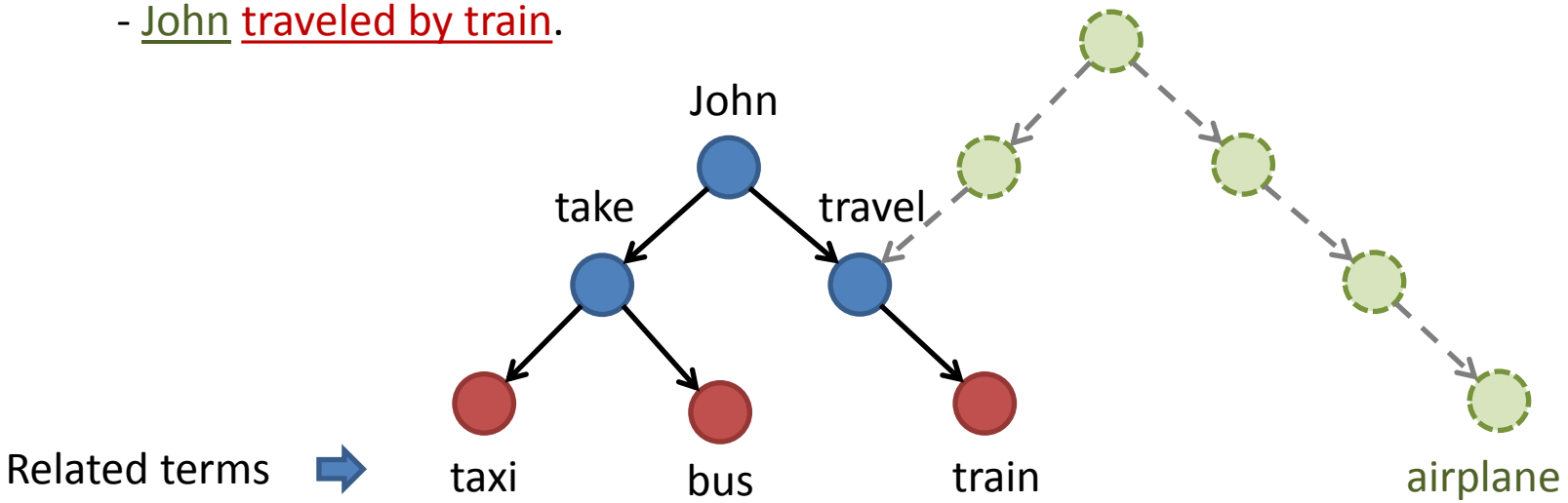
Exemplo:
- John took the taxi.
- Mary took the bus to go home.
- John traveled by train.



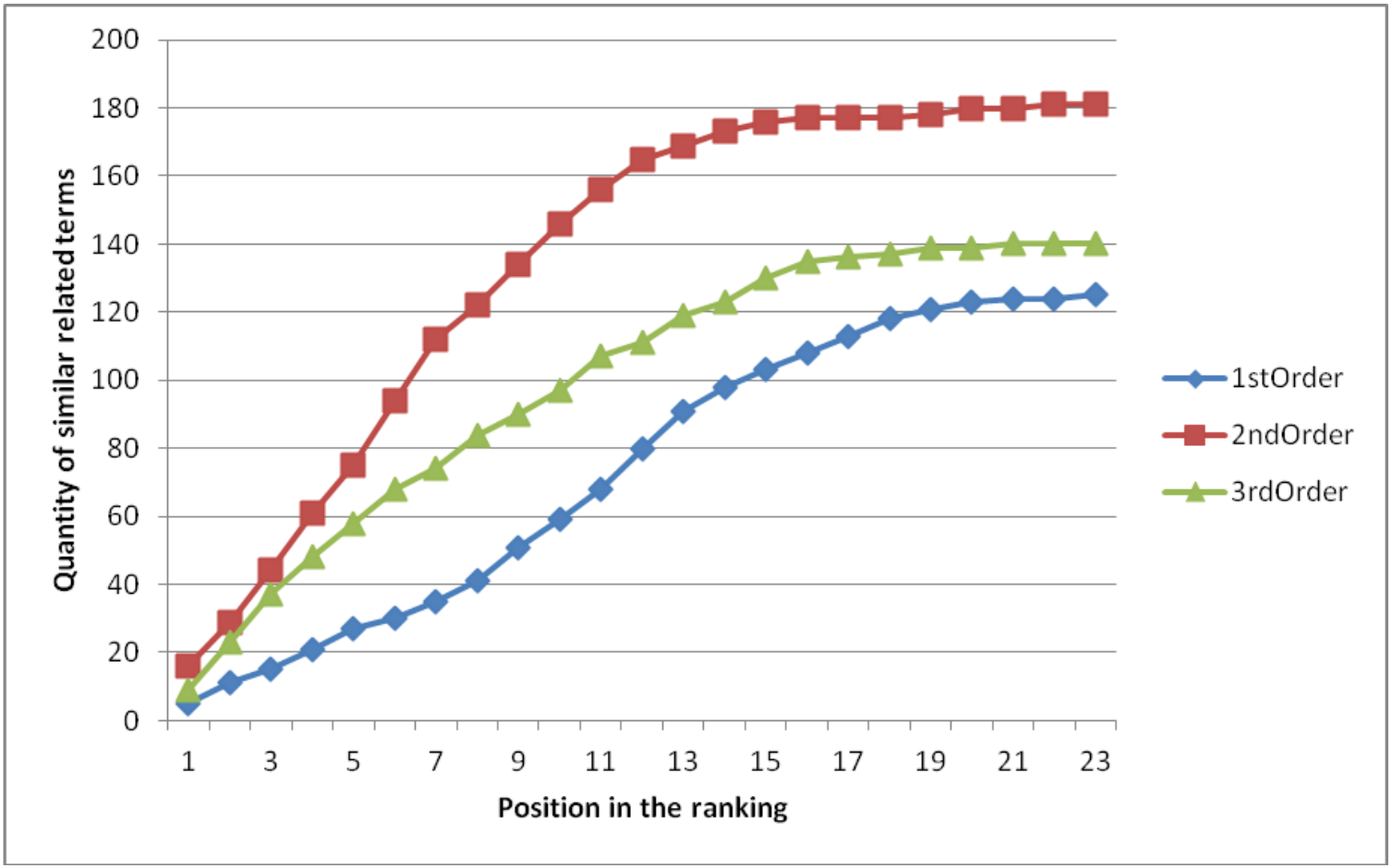
How can I find related terms?

- Using Latent Semantic Analysis (LSA)
  - Third (or more) order co-occurrence

Exemplo:
- John took the taxi.
- Mary took the bus to go home.
- John traveled by train.



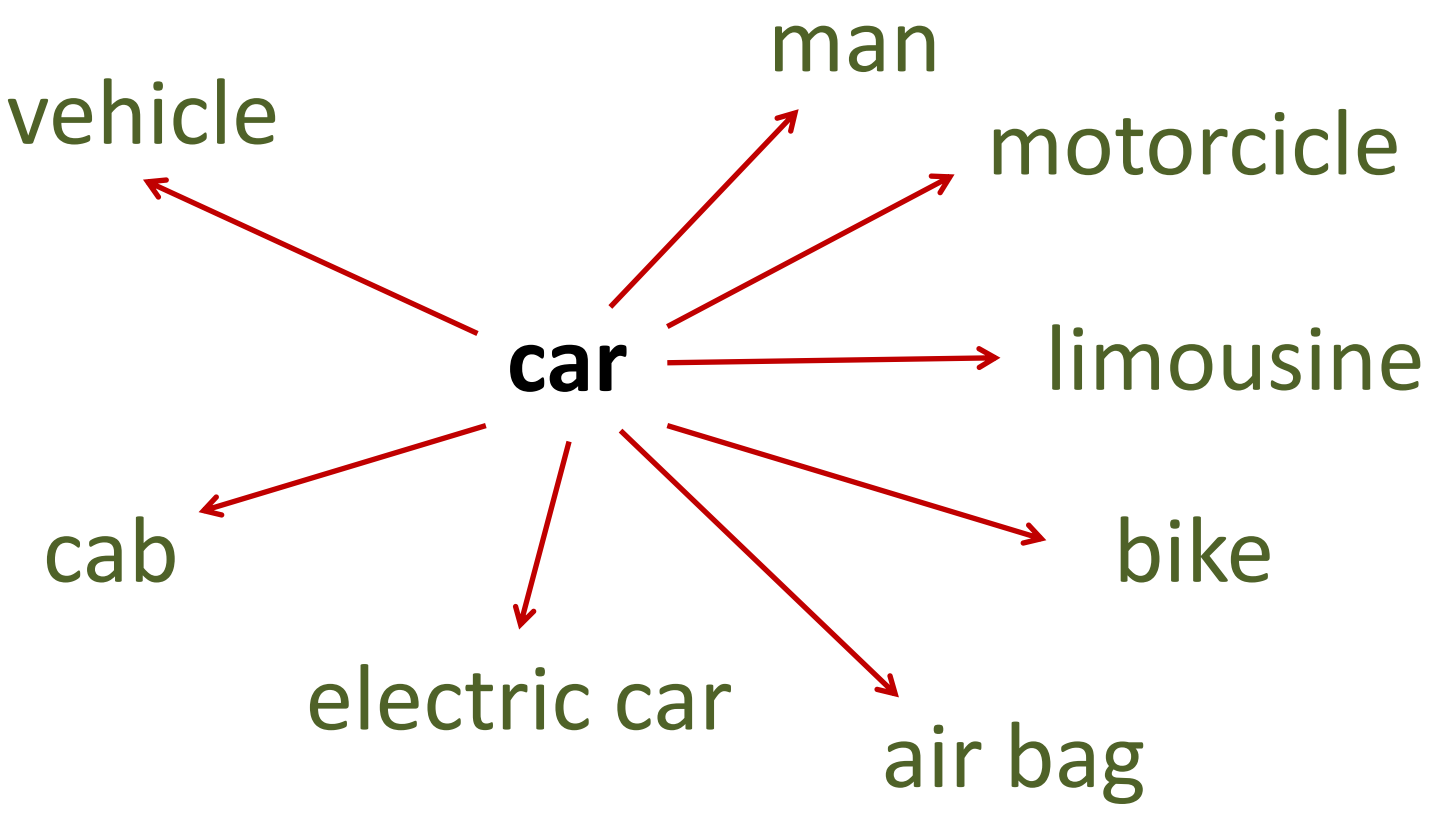
A comparison



But...

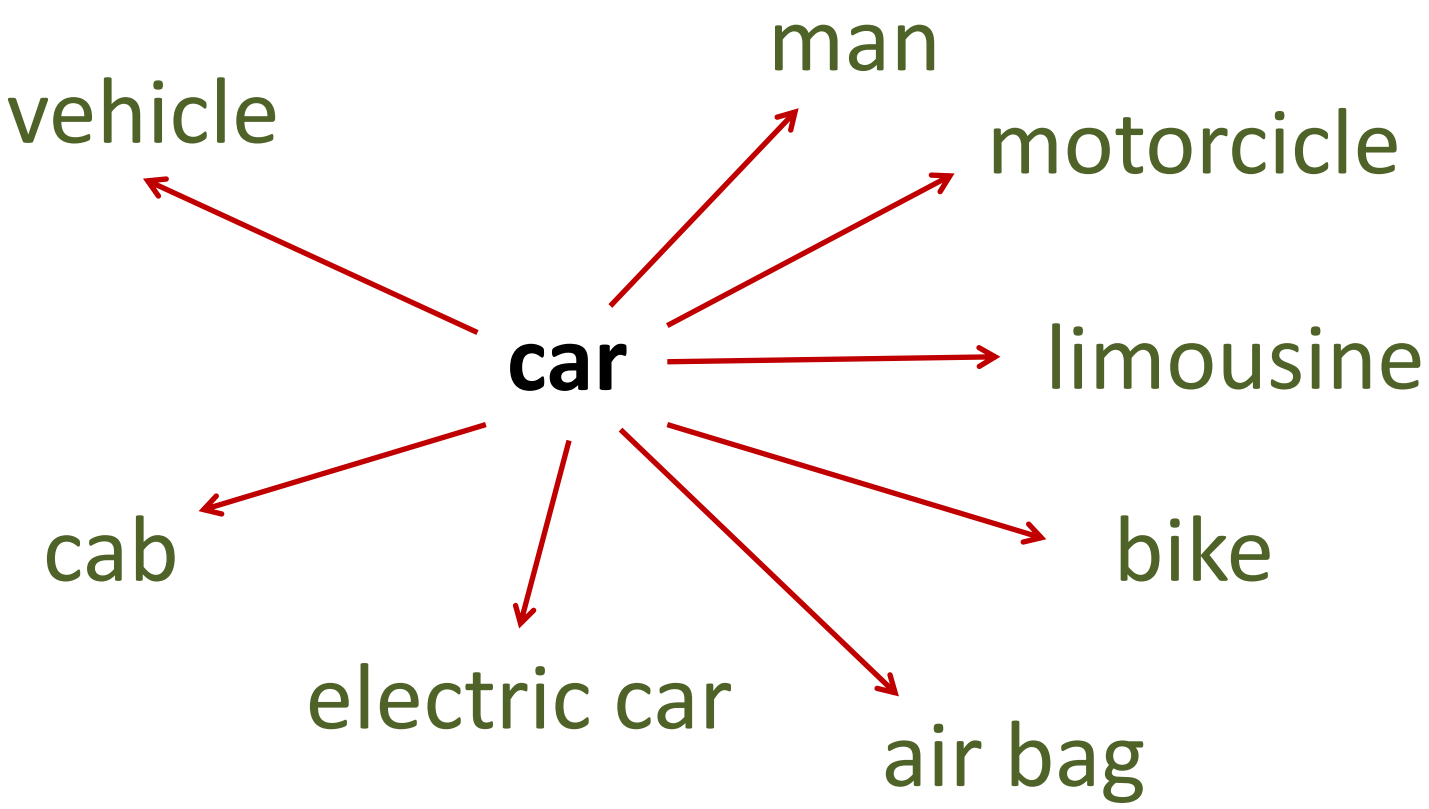
- Cross-lingual LSA (Hassan *et al*. 2012)
  - Appending documents in more than one language

Related terms?



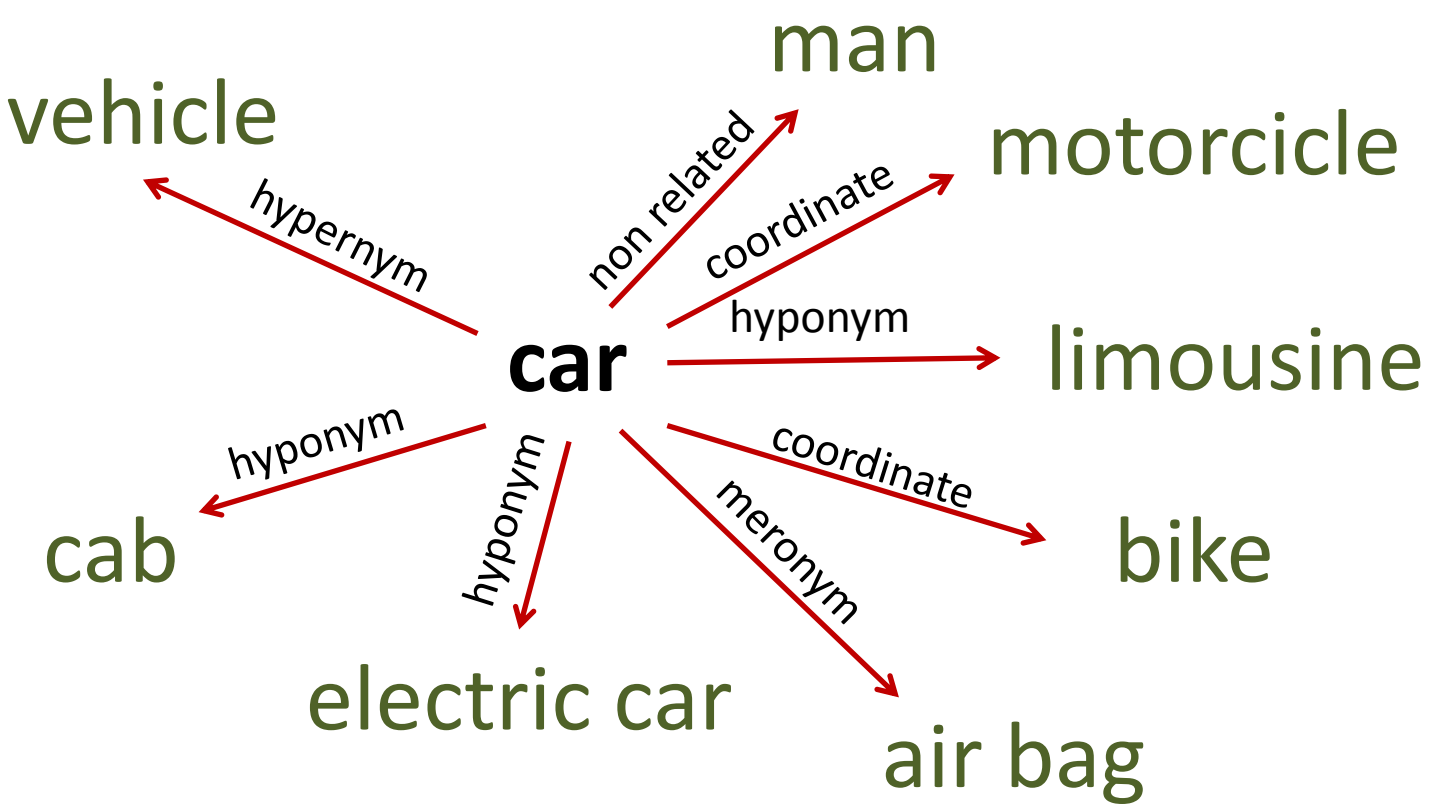
Related terms?

- But what is the relationship?



Related terms?

- But what is the relationship?



3rd CAMELEON Workshop

Porto Alegre, 20 December 2012

Thanks for your attention!

Roger Leitzke Granada

roger.granada@acad.pucrs.br

Pontifical Catholic University of Rio Grande do Sul
Doctoral Program in Computer Science

References

K.W. Church; P. Hanks. “Word association norms, mutual information, and lexicography”. Computational Linguistics, 1990, vol. 16 pp. 22-29.

J. Firth. "A Synopsis of Linguistic Theory 1930-1955". In: Studies in Linguistic Analysis, 1957.

Grefenstette, G. “Explorations in automatic thesaurus discovery”. Kluwer Academic Publishers Norwell, 1994, 306 p.

Harris, Z. (1954). Distributional structure. *Word*, 10(23): 146-162.

T.K. Landauer; S.T. Dumais. “A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge”. Psychological review, 1997, vol. 104, n. 2, pp. 211-240.

D. Lin. "Automatic retrieval and clustering of similar words". In: Proceedings of the 17th international conference on Computational linguistics. 1998, pp. 768-774.

S. Hassan, C. Banea and R. Mihalcea. "Measuring Semantic Relatedness using Multilingual Representations". *SEM 2012: The First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), 2012, 20-29.