FROM RESEARCH TO INDUSTRY

cea tech

list

Semantic Clustering of Relations between Named Entities

PolTAL 2014

Wei Wang, Romaric Besançon, Olivier Ferret

CEA LIST

Vision and Content Engineering Laboratory

Brigitte Grau

LIMSI CNRS

INSTITUT CARNOT CEA LIST digiteo

CNRS LIMSI

Unsupervised Information Extraction

- **Problem**
  - *input*: documents + entities or named entity types
    - No specified type of relations

      ≠ standard information extraction
  - *output*: structured summary of relations between the considered entities

What are the relations between Microsoft and other organizations ?

Unsupervised Information Extraction

- **Microsoft —(?)— <ORG>**

Last year , **Microsoft** purchased **Lookout Software** , a Silicon Valley shareware company operated by two people.

Last March , **Microsoft** acquired **Groove Networks** , a collaboration software company in Beverly , in a $ 120 million deal that brought Groove founder Ray Ozzie to Redmond.

Microsoft launches antitrust complaint against **Google**.

Under the arrangement, which was formalized in the last week, **Microsoft** will cooperate with **Yahoo** on display Ad sales.

… <ORG> purchase <ORG> …
… <ORG> acquire <ORG> …

… <ORG> complaint against <ORG> …

… <ORG> cooperate with <ORG> …

Unsupervised information extraction

- **Two dimensions**
  - extract relations from texts
  - Characterize the relation types by clustering the relations

- **Focus of this work**
  - Large-scale processing
    - > 100 000 relations
  - Open domain → capacity to group relations with varied expressions
    - Semantic criteria

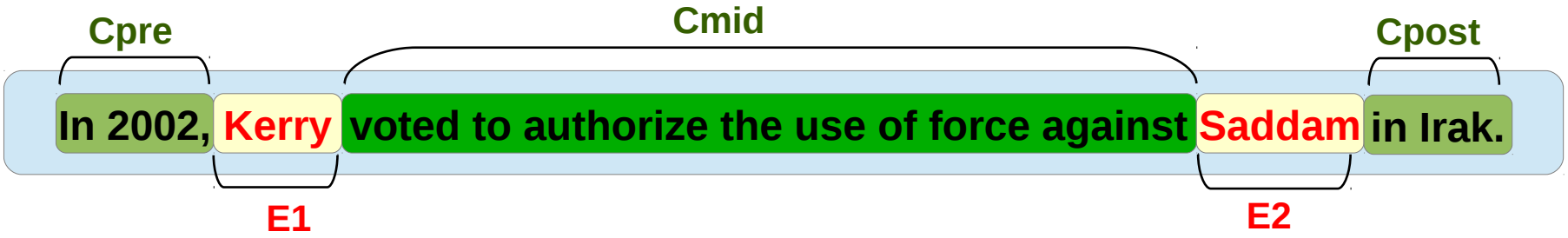
Definition of the relations

- **Characterization**
  - In a sentence
  - Between two arguments
  - Explicitly related to a relation mention

In 2002, Kerry voted to authorize the use of force against Saddam in Irak.

Definition of the relations

- **Characterization**
  - In a sentence
  - Between two arguments
  - Explicitly related to a relation mention



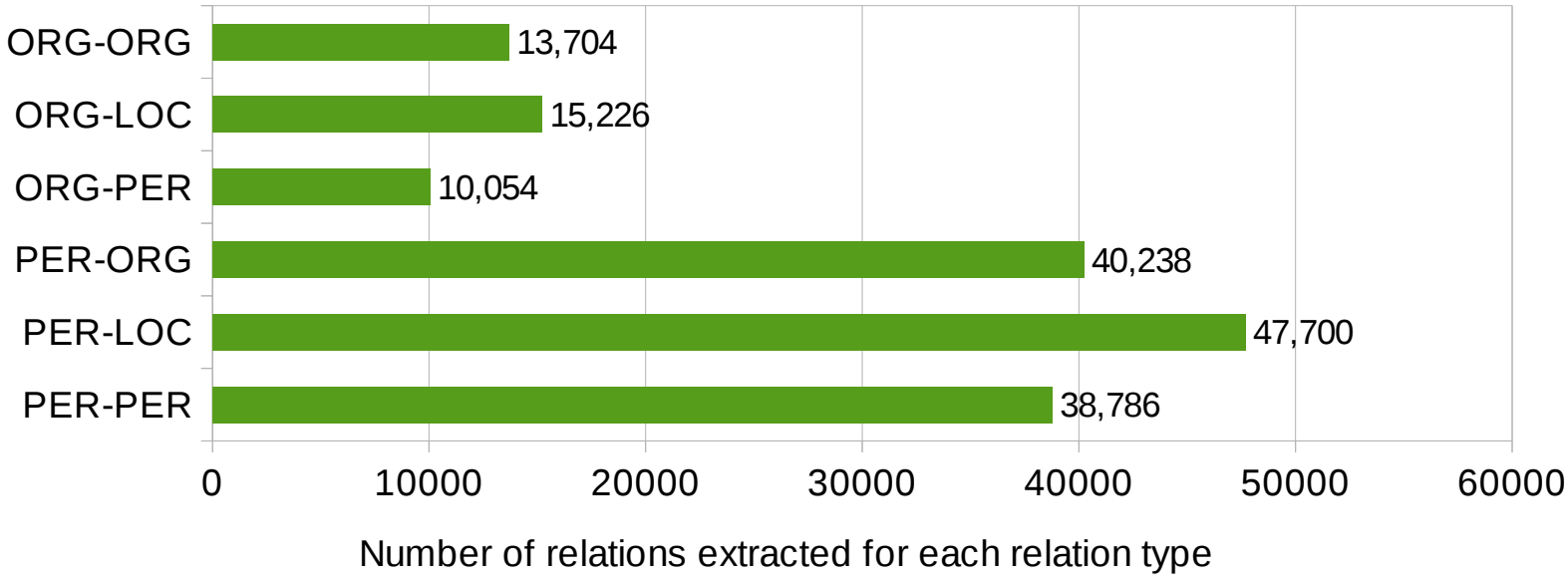
- **Semi-structured form**
  - arguments : named entity pair (E1, E2)
    - Persons, locations, organizations
  - mention : linguistic form of the considered relation
    - 3 parts : before E1 (*Cpre*), between E1 and E2 (***Cmid***), after E2 (*Cpost*)

Relation Extraction : approach

- **Principle**
  - Two-steps approach for « large scale » data
    - Extract candidate relations, based on simple criteria
    - Filter extracted relations with more sophisticated processing
- **Details**
  - Linguistic preprocessing
    - Named Entity Recognition, normalization (LIMA)
  - Simple candidate relation extraction
    - Sentences with 2 NE + verb in-between
  - Filter candidate relations
    - Heuristic filtering : size + complexity of Cmid, reported speech
    - Statistic filtering : linear CRF
      - Non-lexicalized features : POS, POS bigrams…
      - precision = 76,2% ; recall = 78,2%

Relation Extraction : results

- **Experiments**
  - corpus : sub-part of AQUAINT-2 corpus
    18 months of *New York Times* (159 400 documents)
  - filtering : ~ 25% of extracted relations are kept



Relation clustering : problem

- **Goal**
  - Cluster semantically similar relations
- **Two aspects to deal with**
  - Large number of relations : up to ~ 50 000 for one entity type pair
  - Variation of the expression of the relations ← open domain
    - ex. : relation *create* for PERS — ORG

Syntactic variations

create

create the first

that **create**

who **create** the

…

Semantic variations

create **establish**

found

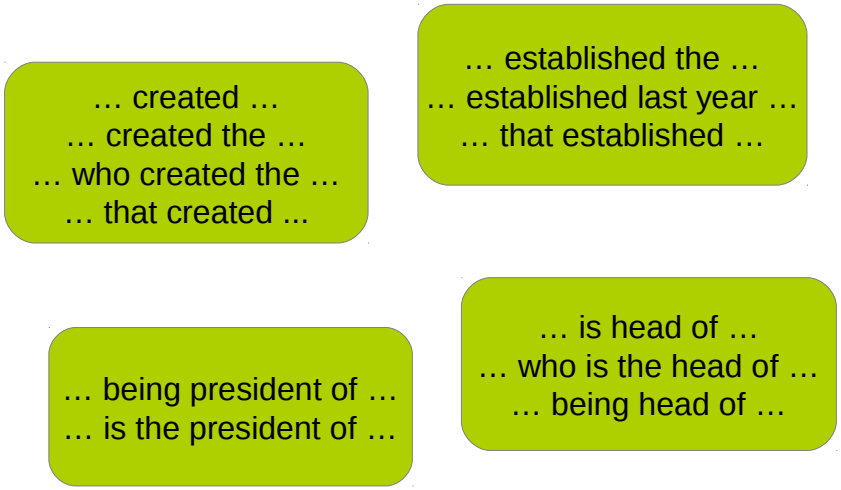
launch

inaugurate

…

Multi-level clustering

Initial clustering

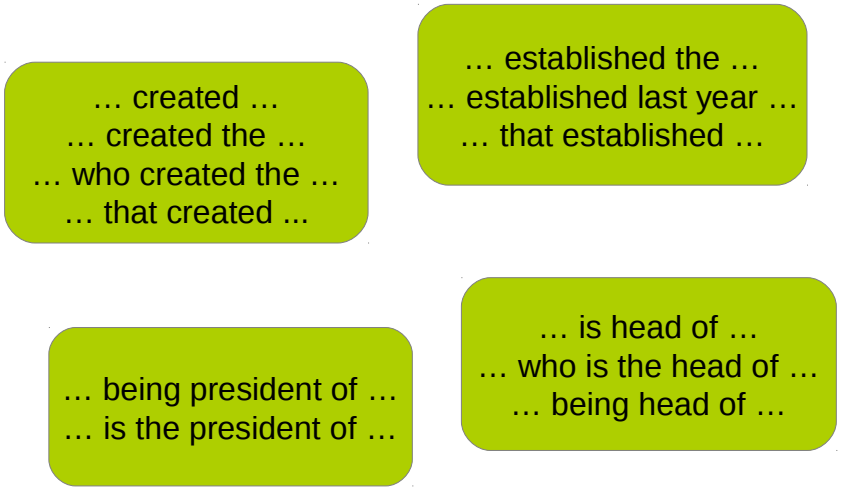


Cluster relations that are similar on their linguistic form to form precise clusters

→ **initial clusters**

Multi-level clustering

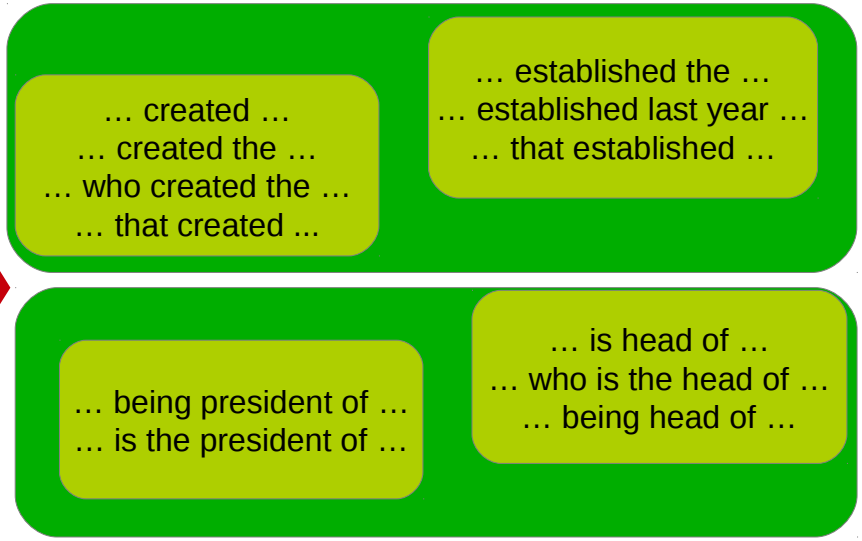
Initial clustering



Cluster relations that are similar on their linguistic form to form precise clusters

→ **initial clusters**

Semantic clustering

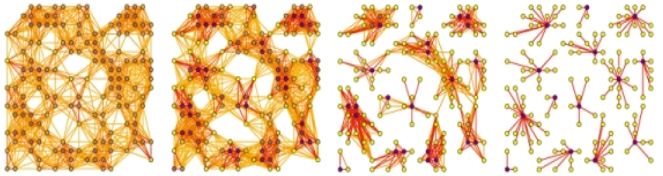


Merge initial clusters that are semantically equivalent to form larger clusters

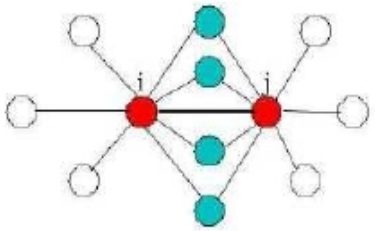
→ **semantic clusters**

Clustering algorithms

- **Global constraints**
  - Large-scale clustering
  - No *a priori* on the number of clusters
- **Initial clustering**
  - Markov Clustering (van Dongen, 2000)
    - Random walk in similarity graph
    - Associated with All Pairs Similarity Search (Bayardo et al., 2007) for efficient computation of similarity graph, using minimum threshold on similarities



- **Semantic clustering**
  - Shared Nearest Neighbor (SNN) Clustering (Ertöz et al., 2002)
    - similarity → number of shared neighbors
      - Allows to consider different similarity measures



Initial clustering

- **Similarities between relations**
  - Globally : Information Retrieval (IR) kind of approach
  - representation : bag-of-words (*Cmid*), using lemmas
    - No term filtering
  - Similarity measure: *Cosinus*
  - 3 weighting schemes for the terms
    - binary : same weighting for all present terms
    - *tf.idf* : relation has the same role as document for IR
    - POS : weight according to part-of-speech

Part-of-speech	weight
Verb, noun, adjective, preposition	*high*
Adverb, pronoun	*medium*
Proper noun	*low*
Symbol, number, determiner	*null*

Initial clustering : evaluation

- **Manually built reference**
  - From the relations extracted from the AQUAINT-2 corpus
  - 80 clusters – 4 420 relations
- **Evaluation Measures**
  - At relation level
    - ∀ couple de relations (R1,R2)

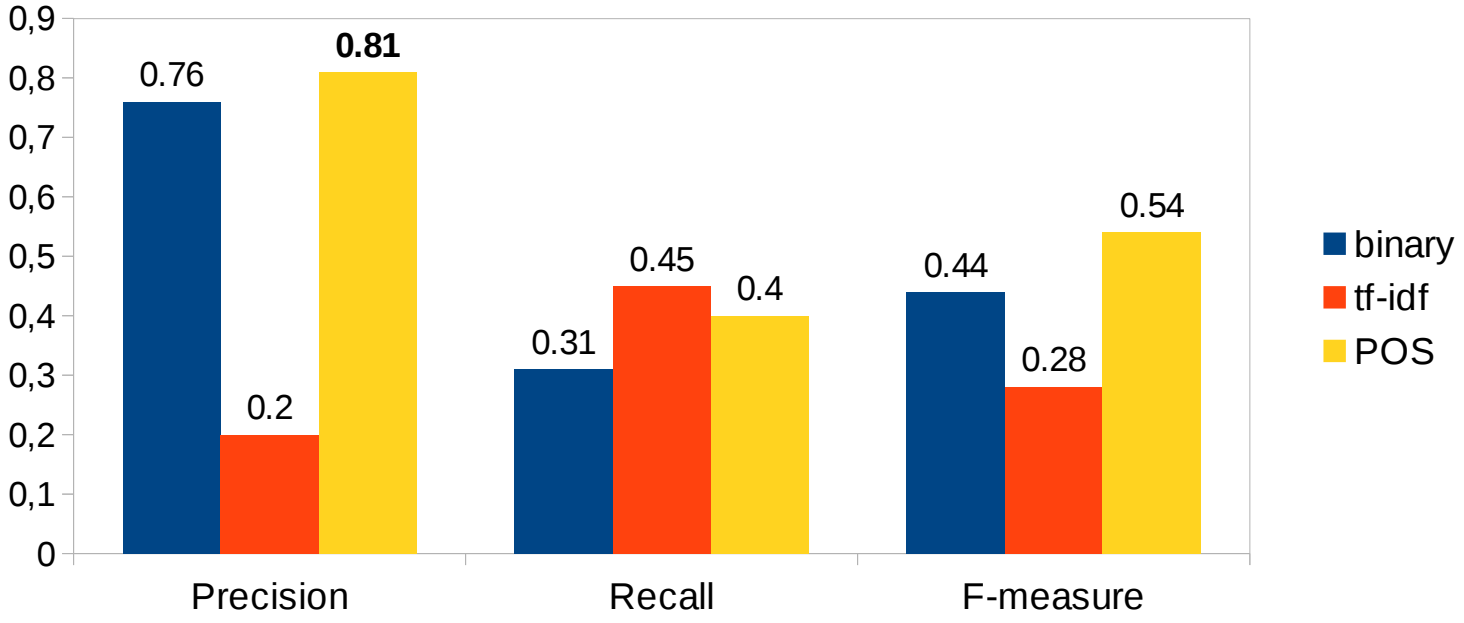
$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{VP}{VP+FN}$$

Result / *Ref*	same cluster	different cluster
same cluster	*TP*	*FN*
different clusters	*FP*	*TN*

  - At cluster level
    - Purity / inverse purity
    - Normalized Mutual Information (NMI)
    - Same tendencies as the ones observed for measures at relation-level

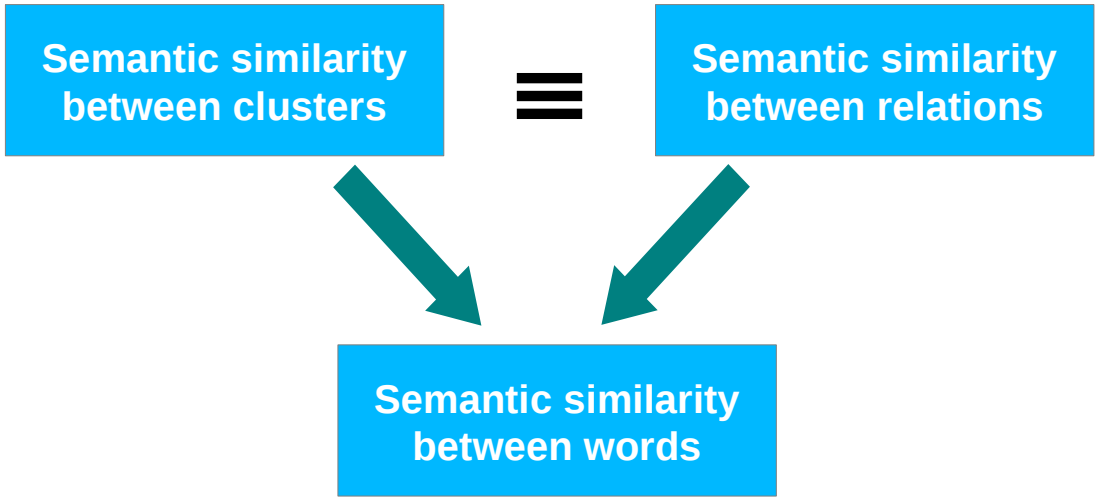
Initial clustering : evaluation



- Low results with *tf.idf* : discriminative words → rare words → not interesting for the relation

Semantic clustering

- **Idea**
  - Clustering of initial clusters
    - ~ 13 600 initial clusters vs. ~ 165 700 relations
  - Parallel between
    - Semantic similarity of basic clusters
    - Semantic similarity of relations

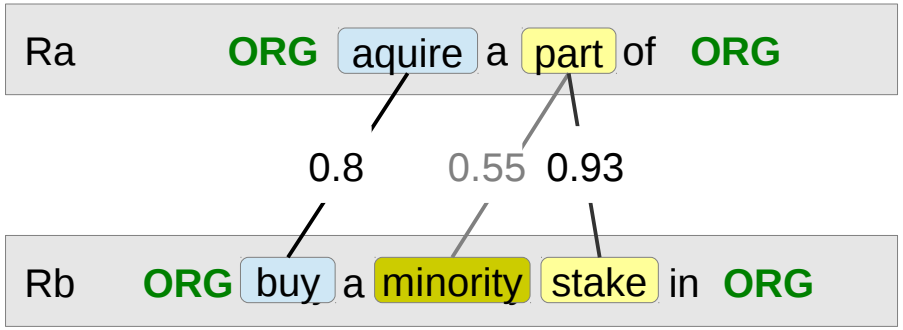


Semantic similarity at word-level

- **2 types of semantic measures → $W_{i,j}$**
  - WordNet-based measures
    - Wu-Palmer (1994)
      - Use the hierarchy of hyperonyms
    - Lin (1998)
      - Add frequency information
  - Distributional measures
    - Use distributional thesaurii built from a corpus
    - 2 types of thesaurii
      - Window-based co-occurrences (Dist-cooc)
      - Syntactic cooccurrences (Dist-syn)
    - thesaurii built using AQUAINT-2 corpus

Semantic similarity at relation-level

- **Same principle as (Mihalcea et al., 2006) for sentence level paraphrases**
  - relation = bag-of-words (*Cmid*), using lemmas
  - For a pair of relations (R_a, R_b)
    - Match each word of R_a with the word of R_b that has the greater similarity



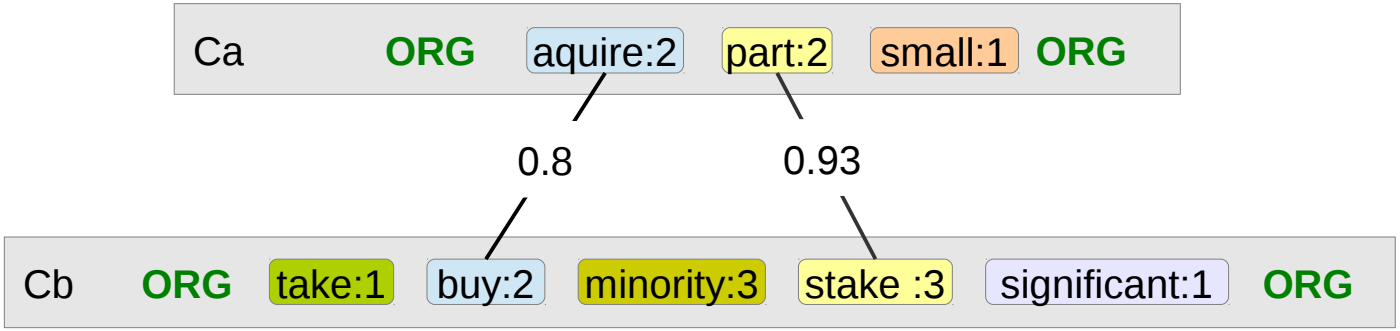
$$\mathrm{sim}(R_a, R_b) = \frac{1}{\sum_{i \in R_a} f_i} \sum_{i \in R_a} \max_{j \in R_b} \{W_{i,j}\} \cdot f_i$$

  - Same for R_b (symmetry)

$$\mathbf{SR_{a,b} = 1/2\ [sim(R_a, R_b) + sim(R_b, R_a)]}$$

Semantic similarity at cluster-level

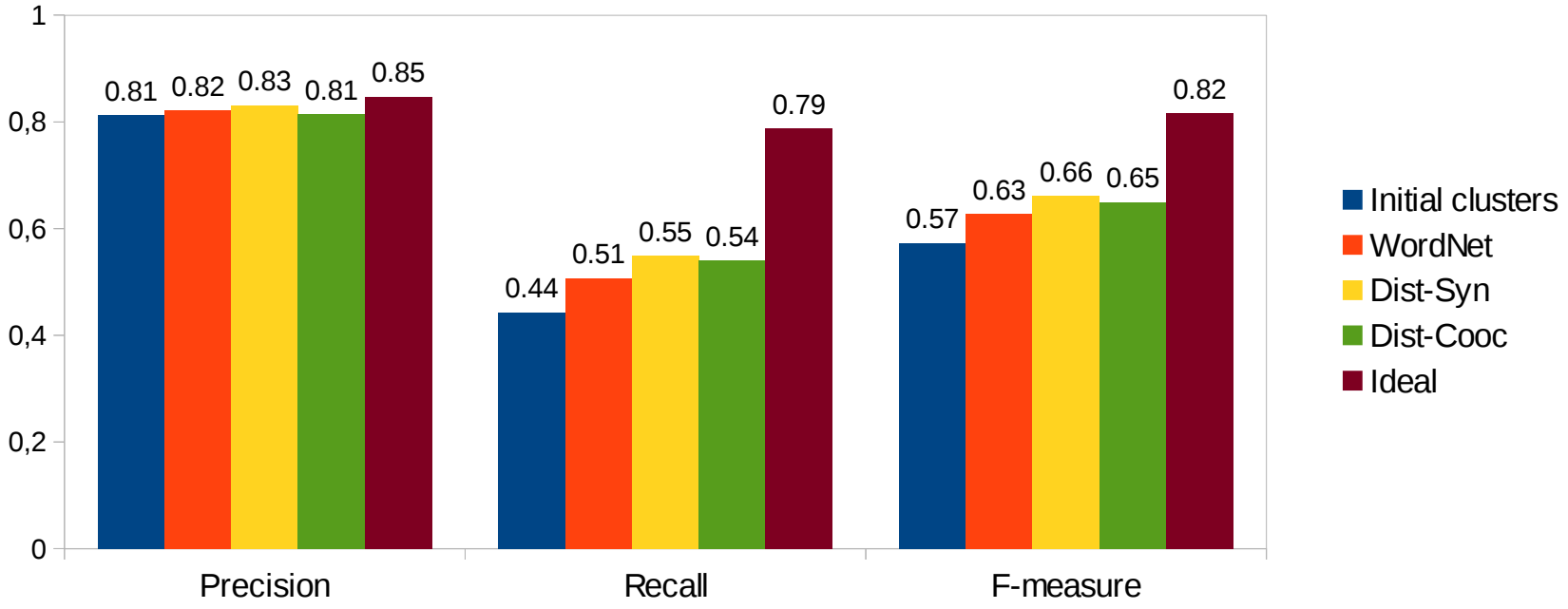
- **Same idea as similarity at relation-level**
  - representation : merge the relation representation → bag-of-words of all lemmas of *Cmid* parts of the relations of the cluster
    - Same matching strategy



    - Same measure (symmetric), with corrective factor for the frequencies

$$SC_{a,b} = \frac{1}{2}\left[\overbrace{\frac{1}{\sum_{\substack{i \in C_a \\ j \in C_b}} f_i \cdot f_j} \sum_{i \in C_a} \max_{j \in C_b} W_{i,j} \cdot f_i \cdot f_j}^{\text{from } C_a} + \overbrace{\frac{1}{\sum_{\substack{i \in C_a \\ j \in C_b}} f_i f_j} \sum_{j \in C_b} \max_{i \in C_a} W_{i,j} \cdot f_i \cdot f_j}^{\text{from } C_b} \right]$$

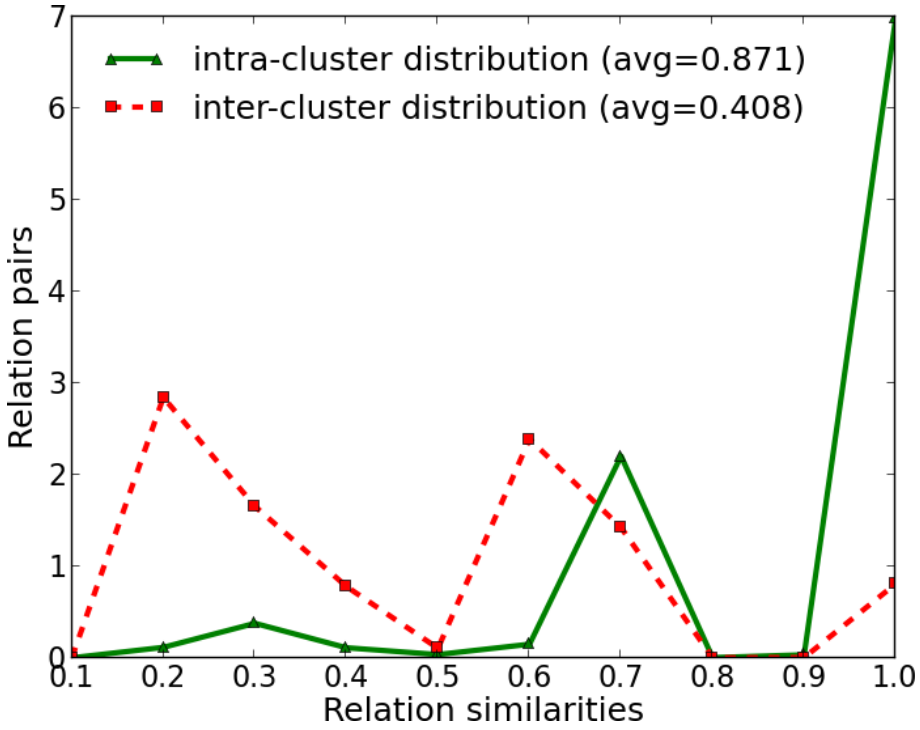
Semantic clustering – Evaluation

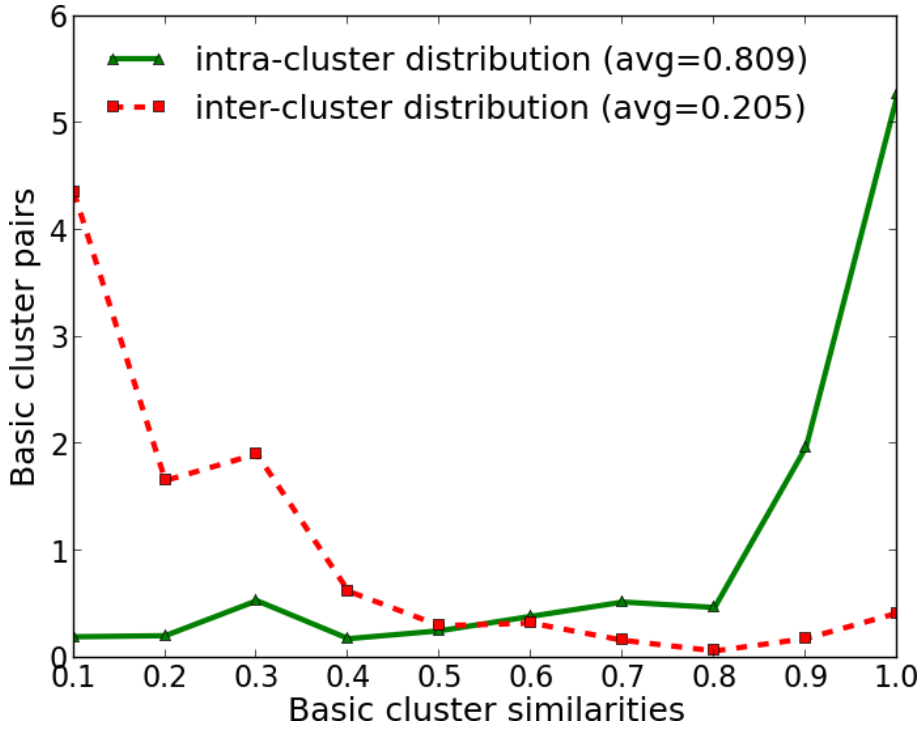


- Clear improvement of recall with semantic clustering, without loss of precision
- Distributional measures tend to be better than WordNet-based measures

Impact of multi-level clustering

- **Similarities between relations vs. Similarities between initial clusters**
  - Initial clusters are better separated than relations





Example of results from semantic clustering

ORG – ORG	*purchase*, buy, acquire, trade, own, *be purchased by*
ORG – LOC	start in, inaugurate service to, open in, initiate flights to
ORG – PER	sign, hire, employ, *interview*, rehire, receive, affiliate
PER – ORG	take over, take control of
PER – LOC	**grab gold in, win the race at**, reign
PER – PER	win over, defeat, beat, topple, defend

Conclusion & Perspectives

- **To sum up**
  - Unsupervised extraction and clustering of relations between named entities
  - Multi-level strategy to deal with
    - Large-scale data
    - Great variability of linguistic forms of the relations
      - Semantic similarity of relations
  - Interest of using distributional resources for semantic relation matching

- **Perspectives**
  - Add a topical clustering + organization w.r.t semantic clustering
  - Extend the relations
    - Cpre, Cpost ; entities → complex terms ; noun-based relations
  - Relation-based search engine