Noun Compound Interpretation

By

Girishkumar Ponkiya

Mentor

Mr. Girish Palshikar

TRDDC, Pune

May-July 2014

Motivational Example

- *Our website homepage logo design was finalized by that indian software designer team.*

```
– (ROOT
    (S
      (NP (PRP$ Our) (NN website) (NN homepage) (NN logo) (NN design))
      (VP (VBD was)
        (VP (VBN finalized)
          (PP (IN by)
            (NP (DT that) (JJ indian) (NN software) (NN designer) (NN team)))))
      (. .)))
```

Motivational Example

- *Our **website homepage logo design** was finalized by that indian **software designer team**.*

- poss(design-5, Our-1)
- **nn(design-5, website-2)**
- **nn(design-5, homepage-3)**
- **nn(design-5, logo-4)**
- nsubjpass(finalized-7, design-5)
- auxpass(finalized-7, was-6)
- root(Root-0, finalized-7)
- prep(finalized-7, by-8)
- det(team-13, that-9)
- amod(team-13, indian-10)
- **nn(team-13, software-11)**
- **nn(team-13, designer-12)**
- pobj(by-8, team-13)

Some more examples..

- Simple (***?***)
  - bone marrow
  - web site design
  - internet connection speed test
  - plastic water bottle
- Complicated (***?***)
  - colon cancer tumor suppressor protein

Simplifying complexity

- colon cancer tumor suppressor protein

  [colon cancer] **[** [tumor suppressor] protein**]**

  - [*tumor suppressor protein*] which is implicated in [*colon cancer*]
    - (IN; LOCATION)
  - [*protein*] that acts as [*tumor suppressor*]
    - (IS; AGENT)
  - [*suppressor*] that inhibits [*tumor*(s)]
    - (OF; PURPOSE)
  - [*cancer*] that occurs in [(the) *colon*]
    - (OF; IN; LOCATION)

Corpus Statistics

- 2-4% of the tokens in various corpora are part of noun compounds (Baldwin and Tanaka, 2004)
  - 2.6% in the British National Corpus
  - 3.9% in the Reuters corpus
  - 2.9% in the Mainichi Shimbun Corpus

- 100M-word British National Corpus (BNC)
  - 939K distinct wordforms
  - 256K distinct noun compounds

Introduction

- Noun Compound (NC): “a sequence of two or more nouns”

    e.g. *box juice, computer science department*

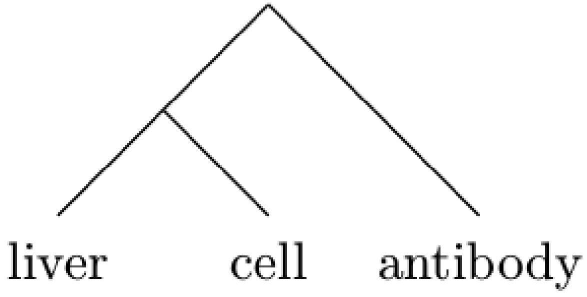
- Individual nouns in the NC are known as “*components*”
- Three main problems:
  - Identifying noun compound
  - Syntactic analysis (*bracketing*)
  - Semantic Relation assignment

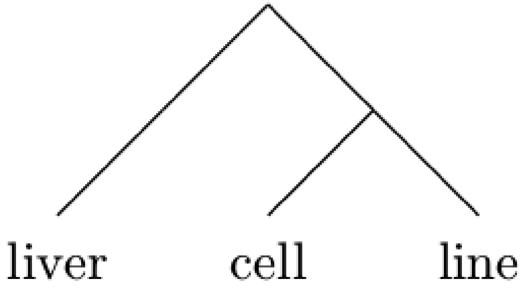
Bracketing

- Determining syntactic structure
- Examples:
  - (1) *liver cell antibody*

    [[liver cell] antibody]
  - (2) *liver cell line*

    [liver [cell line]]





Bracketing

- Methods

  e.g. *computer science department*, *linguistics graduate program*

  - **Adjacency model**

    based on frequency of (N1,N2) and (N2,N3) in bia-gram data

  - **Dependency model**

    based on frequency of (N1,N3) and (N2,N3) in dependecy data

  - **Hybrid**
    - n-gram, adjacency, dependecy, and some more features

Semantic Interpretation

- Approaches
  - Rule based (Vanderwende, 1994)
  - Statistical
    - Analogy based resoning
      - “similar component words should have the same SR”
        e.g. *cat:meow* <=> *dog:bark*
    - semantic disambiguation
      - Disambiguation relative to an underlying predicate or paraphrase

Levi's Theory (1978)

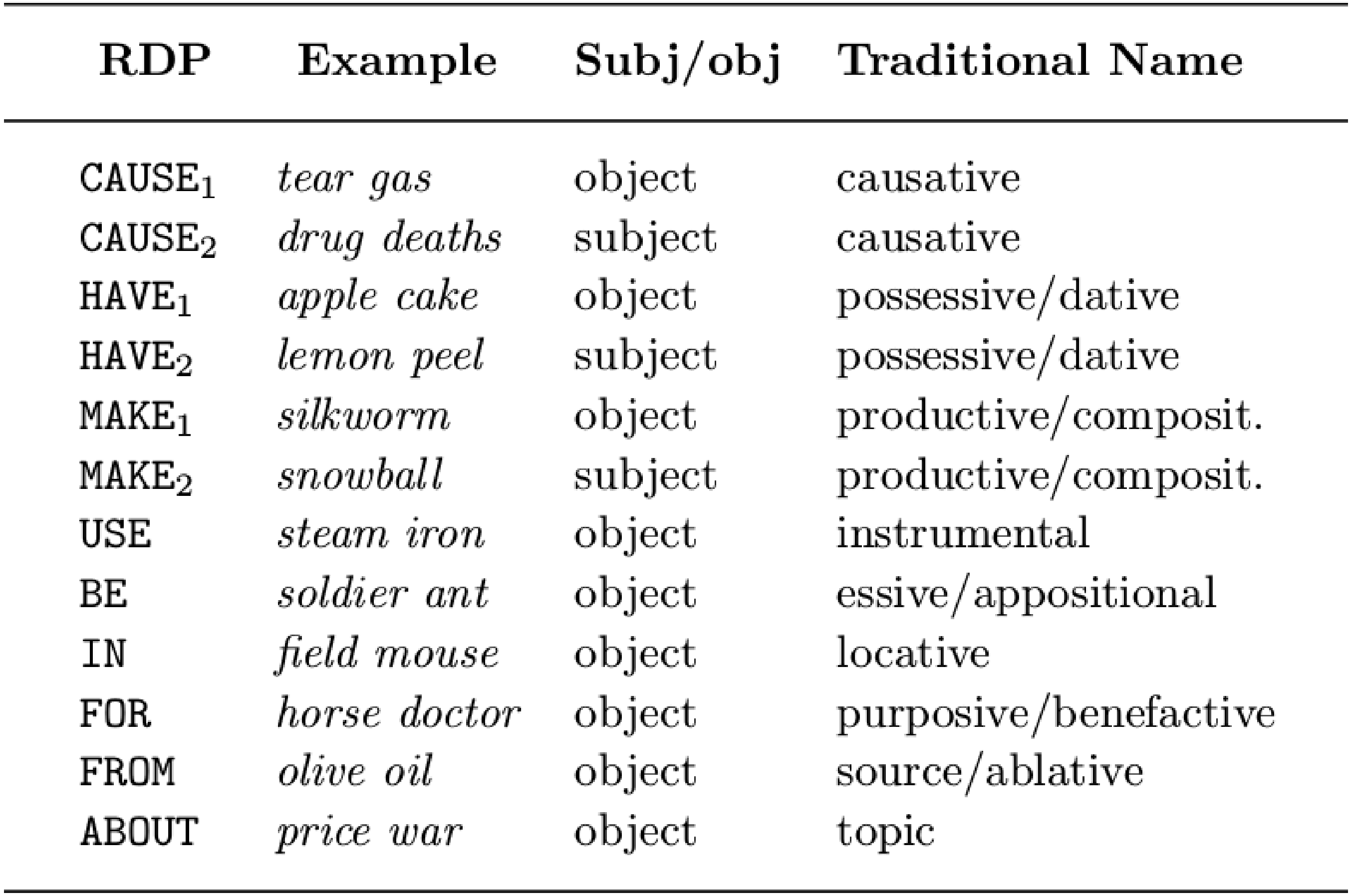
- Idea: study how noun compound can be derived
- Two syntactic processes:
  - predicate nominalization
    - For example, in sentence:

      ..the President refused General MacArthur’s request..

      → *presidential refusal*
  - predicate deletion
    - Example:

      pie made of apples → *apple pie*
    - Proposed set of abstract recoverably deletable predicates

Recoverably Deletable Predicates



RDP	**Example**	**Subj/obj**	**Traditional Name**
CAUSE$_1$	*tear gas*	object	causative
CAUSE$_2$	*drug deaths*	subject	causative
HAVE$_1$	*apple cake*	object	possessive/dative
HAVE$_2$	*lemon peel*	subject	possessive/dative
MAKE$_1$	*silkworm*	object	productive/composit.
MAKE$_2$	*snowball*	subject	productive/composit.
USE	*steam iron*	object	instrumental
BE	*soldier ant*	object	essive/appositional
IN	*field mouse*	object	locative
FOR	*horse doctor*	object	purposive/benefactive
FROM	*olive oil*	object	source/ablative
ABOUT	*price war*	object	topic

O Seaghdha's Thoery (2007)

- Revised the inventory of Levi (1978)
  - The inventory of relations should have good **coverage**
    - *history teacher, woman driver*
  - Relations should be disjunct, and should describe a **coherent** concept
    - Overlapping category boundaries
    - annotation guidelines
  - The **class distribution** should not be overly skewed or sparse
  - The concepts underlying the relations should **generalize** to other linguistic phenomena
  - The guidelines should make the **annotation process** as simple as possible
  - The categories should provide useful semantic information.
- 2000 samples in dataset

Warren's Theory (1978)

- Based on study of Brown corpus
- Abstract semantic relations organized into a four-level hierarchy
  - **CONSTITUTE**: A is something that wholly constitutes B, or vice-versa
    - Source-Result, Result-Source, Copula
  - **POSSESSION**: A is something of which B is a part or a feature or vice versa
    - Part-Whole, Whole-Part, Size-Whole
  - **LOCATION**: A is the location or origin of B (in time or space)
    - Place-OBJ, Time-OBJ, Origin-OBJ
  - **ACTIVITY-ACTOR**: The comment indicates the activity or interest with which B is habitually concerned
  - **RESEMBLANCE**: A indicates something that B resembles
    - Comparant-Compared
  - **PURPOSE**: A is purpose of B, or vice-versa.

Improving Warren's Theory

- Barker & Szpakowicz (1998)
  - Flat 20 relations
  - From Wall Street Journal (Kim and Baldwin, 2005)
    - 2,169 unique 2-term NC
    - 1,571 unique 3-term NC
- Nastase & Szpakowicz (2003)
  - 5 coarse-grained super-relations
  - 30 fine-grained relations
  - 600 samples in dataset

A Lexical Semantic Approach to Interpreting and Bracketing English Noun Compounds

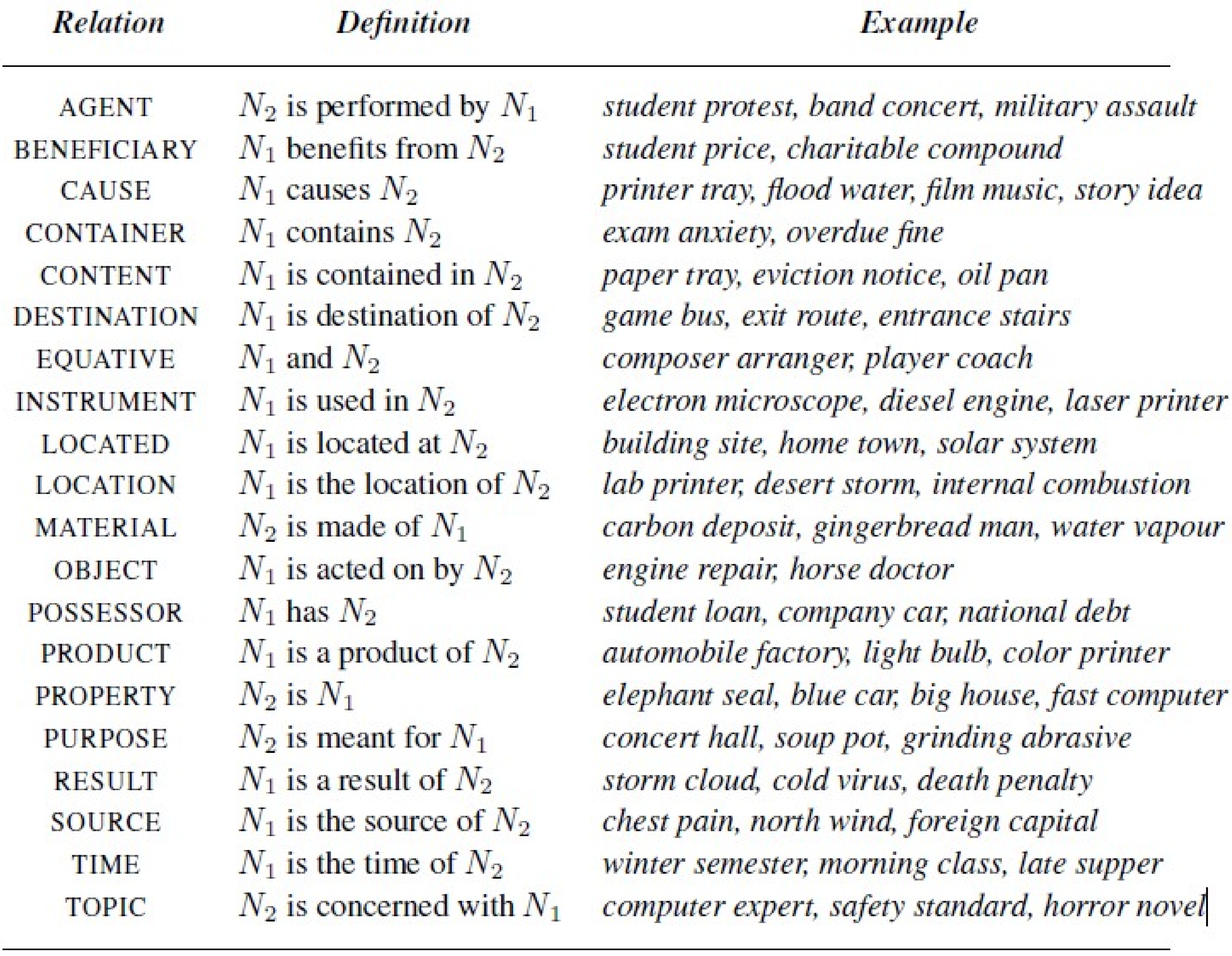
Su Nam Kim and **Timothy Baldwin**

Overview

- Goal
  - Automatic NC interpretation
- Approach
  - Analogical, based on WordNet similarity
- Other
  - NC interpretation helps bracketing

Semantic Relations

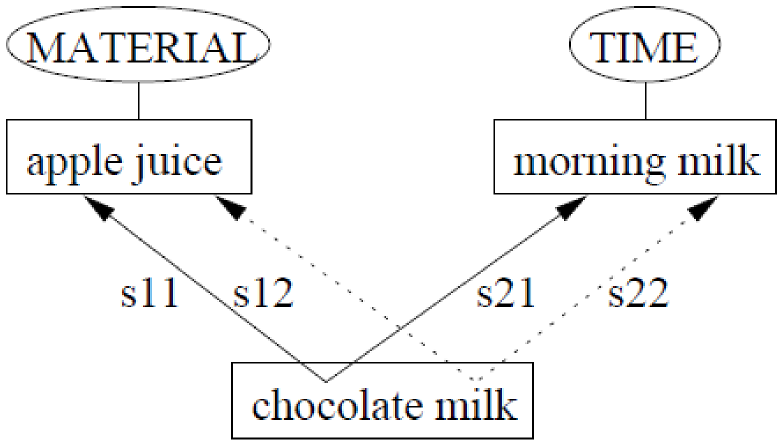
- Used the set of 20 SRs proposed by Barker and Szpakowicz (1998)
  - Relatively well-established in NLP research
  - Found to adequately capture the dataset used in this paper
- List of SRs in next slide



Relation	*Definition*	*Example*
AGENT	N_2 is performed by N_1	*student protest, band concert, military assault*
BENEFICIARY	N_1 benefits from N_2	*student price, charitable compound*
CAUSE	N_1 causes N_2	*printer tray, flood water, film music, story idea*
CONTAINER	N_1 contains N_2	*exam anxiety, overdue fine*
CONTENT	N_1 is contained in N_2	*paper tray, eviction notice, oil pan*
DESTINATION	N_1 is destination of N_2	*game bus, exit route, entrance stairs*
EQUATIVE	N_1 and N_2	*composer arranger, player coach*
INSTRUMENT	N_1 is used in N_2	*electron microscope, diesel engine, laser printer*
LOCATED	N_1 is located at N_2	*building site, home town, solar system*
LOCATION	N_1 is the location of N_2	*lab printer, desert storm, internal combustion*
MATERIAL	N_2 is made of N_1	*carbon deposit, gingerbread man, water vapour*
OBJECT	N_1 is acted on by N_2	*engine repair, horse doctor*
POSSESSOR	N_1 has N_2	*student loan, company car, national debt*
PRODUCT	N_1 is a product of N_2	*automobile factory, light bulb, color printer*
PROPERTY	N_2 is N_1	*elephant seal, blue car, big house, fast computer*
PURPOSE	N_2 is meant for N_1	*concert hall, soup pot, grinding abrasive*
RESULT	N_1 is a result of N_2	*storm cloud, cold virus, death penalty*
SOURCE	N_1 is the source of N_2	*chest pain, north wind, foreign capital*
TIME	N_1 is the time of N_2	*winter semester, morning class, late supper*
TOPIC	N_2 is concerned with N_1	*computer expert, safety standard, horror novel*

NC Interpretation: Approach

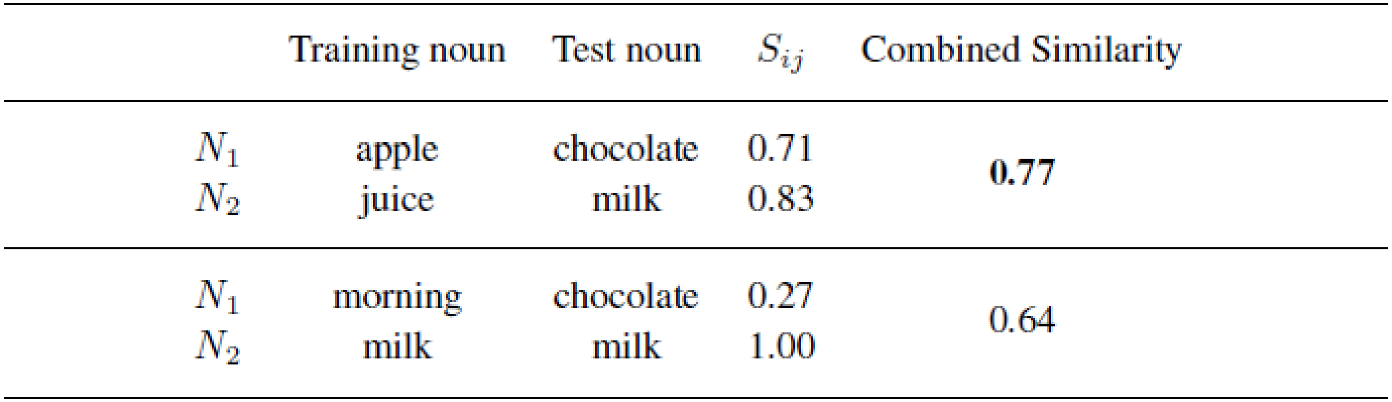
- For 2-term NC



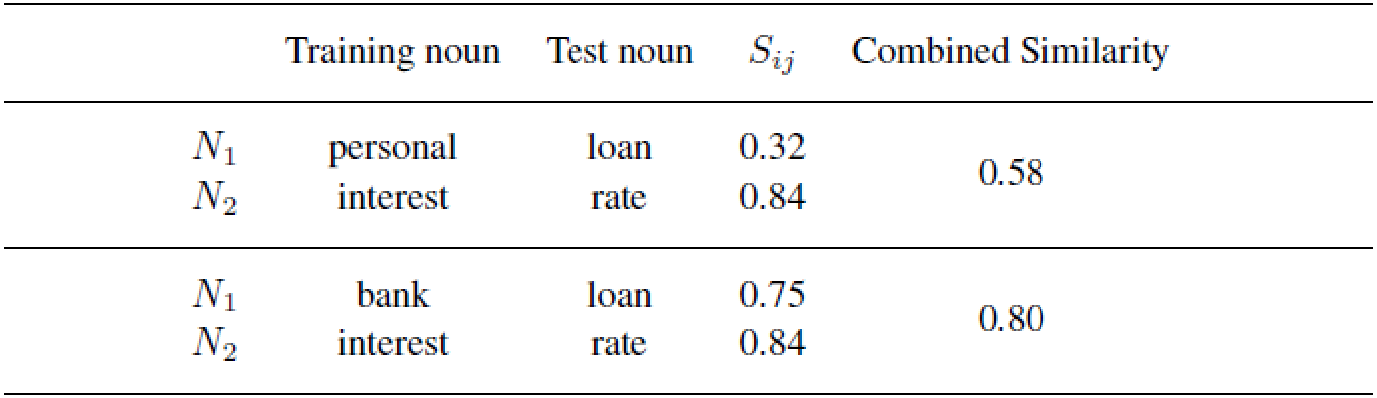
$$S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) = \alpha S1 + (1 - \alpha) S2$$

NC Interpretation: Example

- For 2-term NC



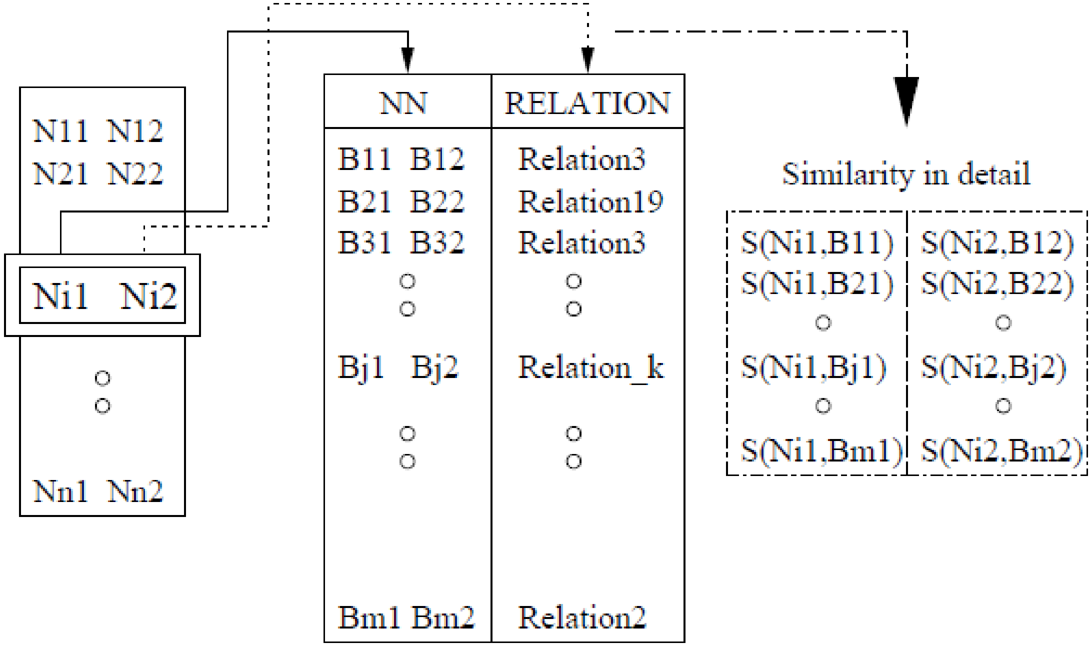
	Training noun	Test noun	S_{ij}	Combined Similarity
N_1	apple	chocolate	0.71	**0.77**
N_2	juice	milk	0.83	
N_1	morning	chocolate	0.27	0.64
N_2	milk	milk	1.00	



	Training noun	Test noun	S_{ij}	Combined Similarity
N_1	personal	loan	0.32	0.58
N_2	interest	rate	0.84	
N_1	bank	loan	0.75	0.80
N_2	interest	rate	0.84	

NC Interpretation: Approach

- For 2-term NC



N11 N12
N21 N22
Ni1 Ni2
○
○
Nn1 Nn2

NN	RELATION
B11 B12	Relation3
B21 B22	Relation19
B31 B32	Relation3
○	○
○	○
Bj1 Bj2	Relation_k
○	○
○	○
Bm1 Bm2	Relation2

Similarity in detail

S(Ni1,B11)	S(Ni2,B12)
S(Ni1,B21)	S(Ni2,B22)
○	○
S(Ni1,Bj1)	S(Ni2,Bj2)
○	○
S(Ni1,Bm1)	S(Ni2,Bm2)

$$m = \operatorname*{argmax}_{j} S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2}))$$

Data Collection

- Source: Wall Street Journal
- Collected 2-term and 3-terms NCs
  - 2,169 unique 2-term NCs
  - 1,571 unique 3-term NCs

Data Annotation

- 2 trained human annotator
- First step: bracketing 3-term NC
- Second step: tagged outermost 2-term NC

  (N2 N3) for ((N1 N2) N3), and

  (N1 N3) for (N1 (N2 N3))
- Multiple SRs were assigned

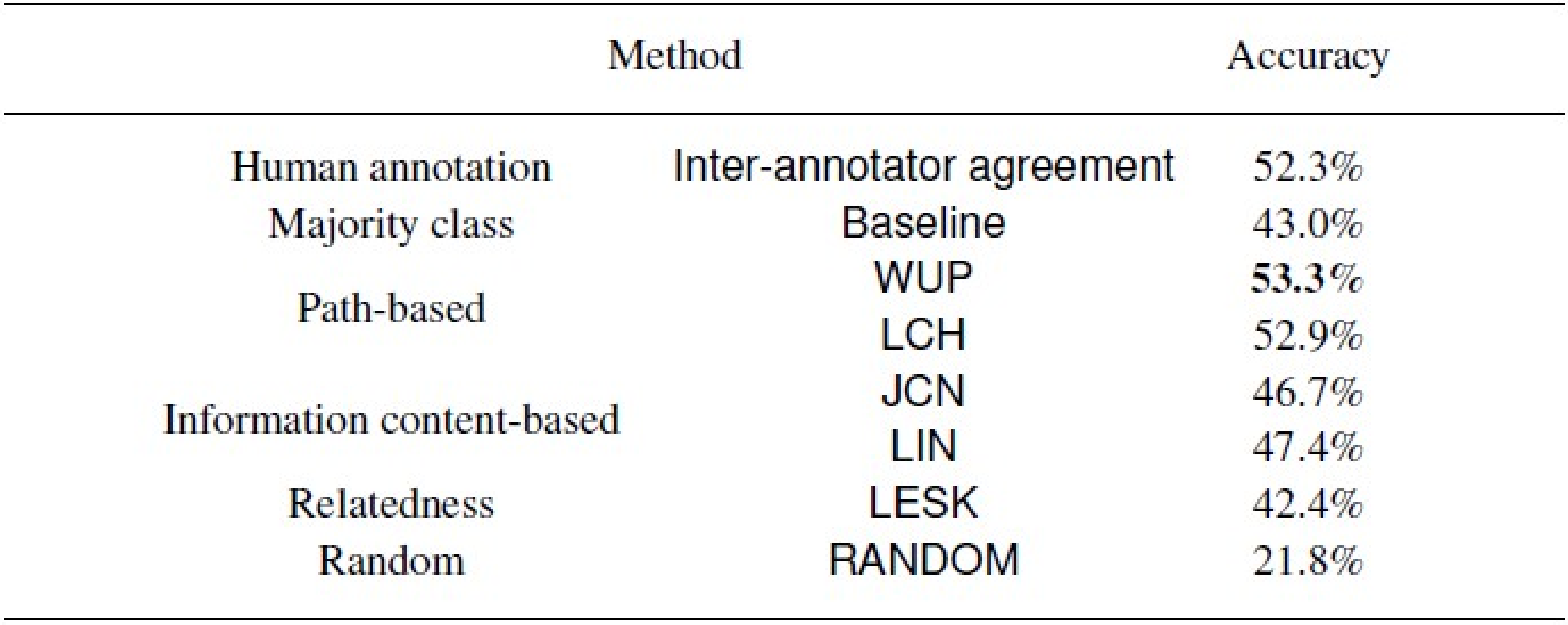
  e.g. *debt cost* : SOURCE or CAUSE ??
- Agreement for SR
  - 2-term: 52.31 %
  - 3-term: 49.28 %

	2-term NCs				*3-term NCs*			
Relation	Test		Training		Test		Training	
	N+	M	N+	M	N+	M	N+	M
AGENT	10	1	5	0	9	0	7	1
BENEFICIARY	10	1	7	1	2	0	3	0
CAUSE	54	5	74	3	21	0	18	0
CONTAINER	13	4	19	3	13	1	7	2
CONTENT	40	2	34	2	23	0	18	0
DESTINATION	1	0	2	0	0	0	1	0
EQUATIVE	9	0	17	1	1	0	2	1
INSTRUMENT	6	0	11	0	2	0	3	0
LOCATED	12	1	16	2	3	0	5	0
LOCATION	29	9	24	4	19	0	27	0
MATERIAL	12	0	14	1	10	0	11	0
OBJECT	88	6	88	5	22	6	26	3
POSSESSOR	33	1	22	1	25	4	21	6
PRODUCT	27	0	32	6	27	1	26	1
PROPERTY	76	3	85	3	33	0	43	0
PURPOSE	159	13	161	9	89	7	95	6
RESULT	7	0	8	0	3	0	4	0
SOURCE	75	11	99	15	61	0	44	1
TIME	25	1	19	0	19	0	24	0
TOPIC	465	24	447	39	438	16	437	15
TOTAL	1163	82	1184	96	820	35	822	36

Experiments #1

- For 2-term NC
- With equal weight for head and modifier similarities
- *k*-NN methods with various *k* values
  - *k*=1 was found better
- Contribution of training-data size

Experiment #1: Result



Method		Accuracy
Human annotation	Inter-annotator agreement	52.3%
Majority class	Baseline	43.0%
Path-based	WUP	**53.3%**
	LCH	52.9%
Information content-based	JCN	46.7%
	LIN	47.4%
Relatedness	LESK	42.4%
Random	RANDOM	21.8%

Table 7. *Accuracy of NC interpretation for the different WordNet-based scoring methods over our 2-term NC dataset*

Experiment #1: Result

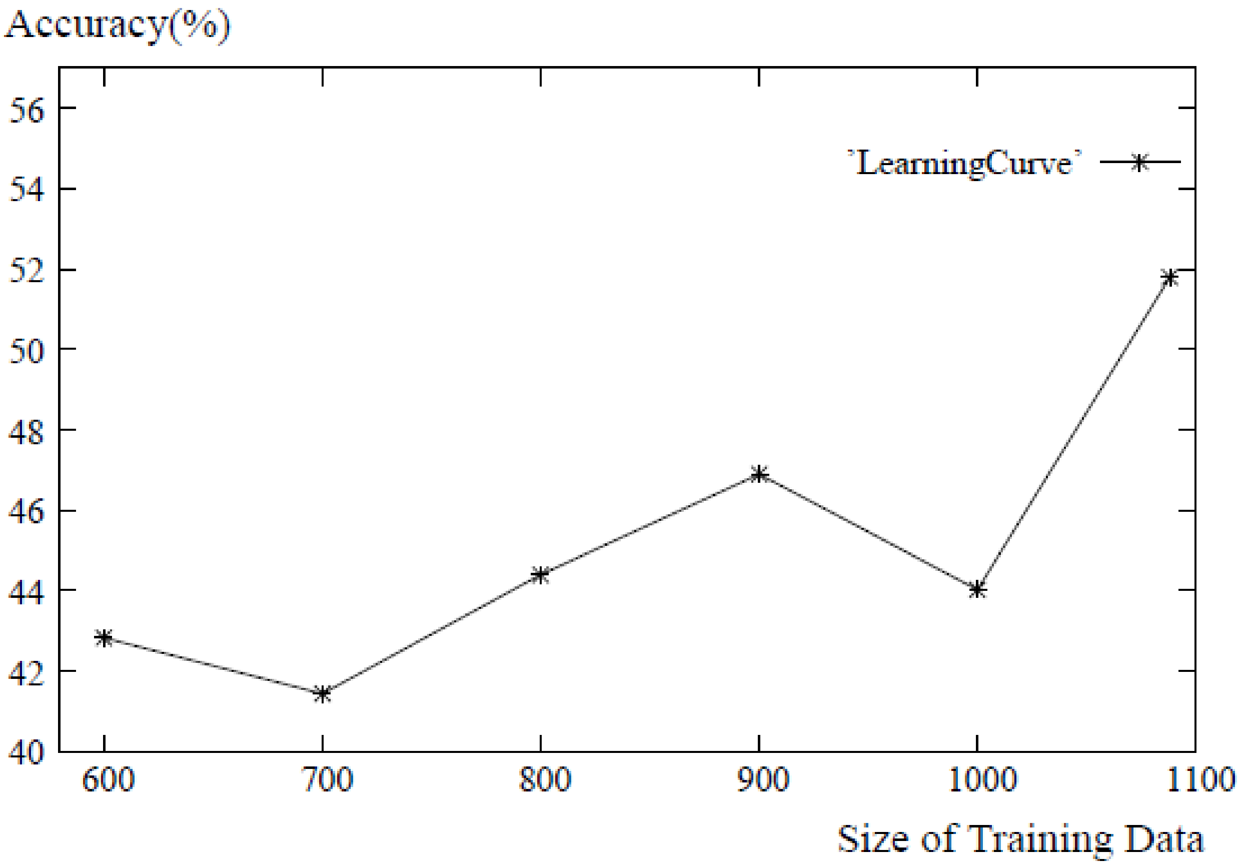


Fig. 3. Learning Curve with respect to the size of the training data

Experiment #2

- To check relative contribution of head and modifier

$$S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2})) = \alpha S1 + (1 - \alpha) S2$$

- For example
  - Head playes important role in PROPERTY relation e.g. *fairy penguin*
  - Modifirer plays important role in TIME relation i.e. *winter coat*

Experiment #2: Result

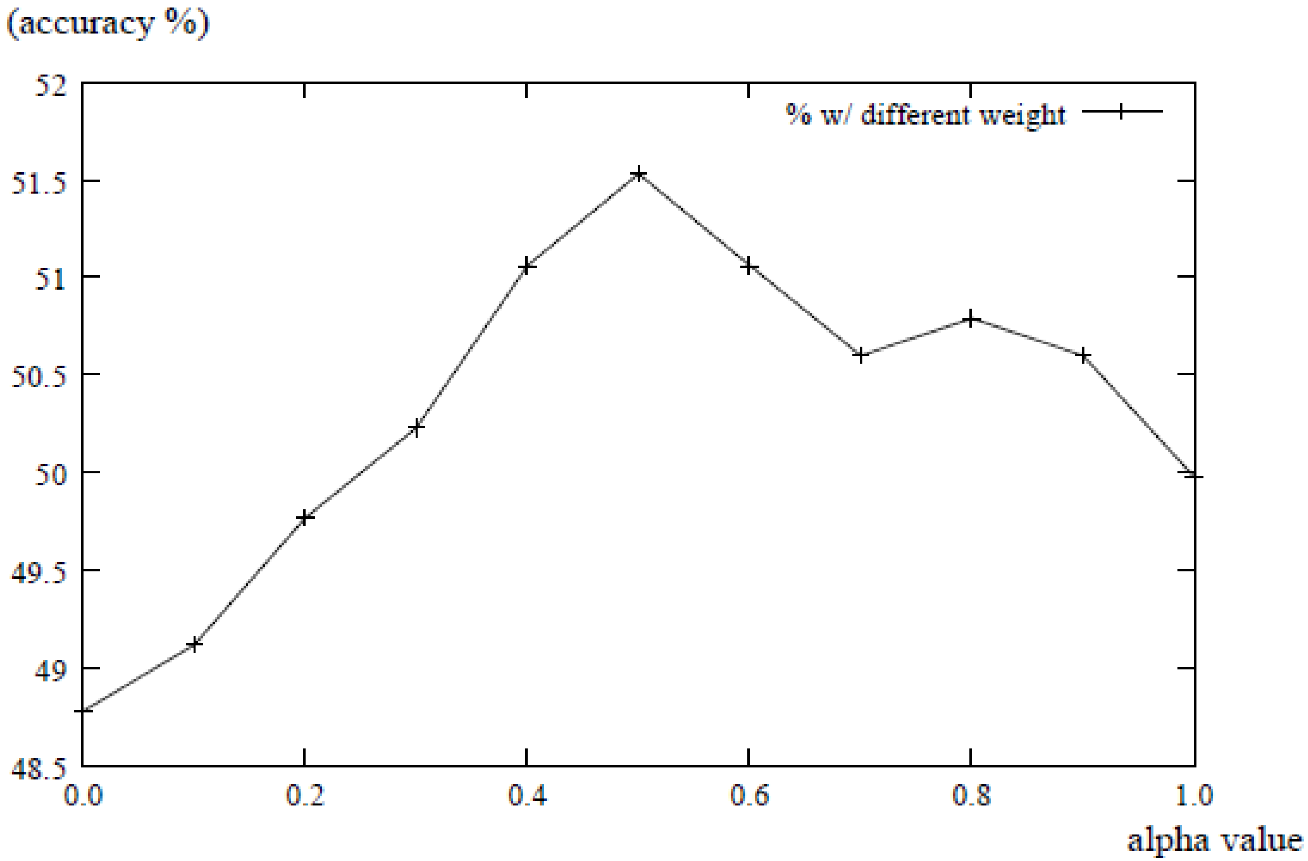
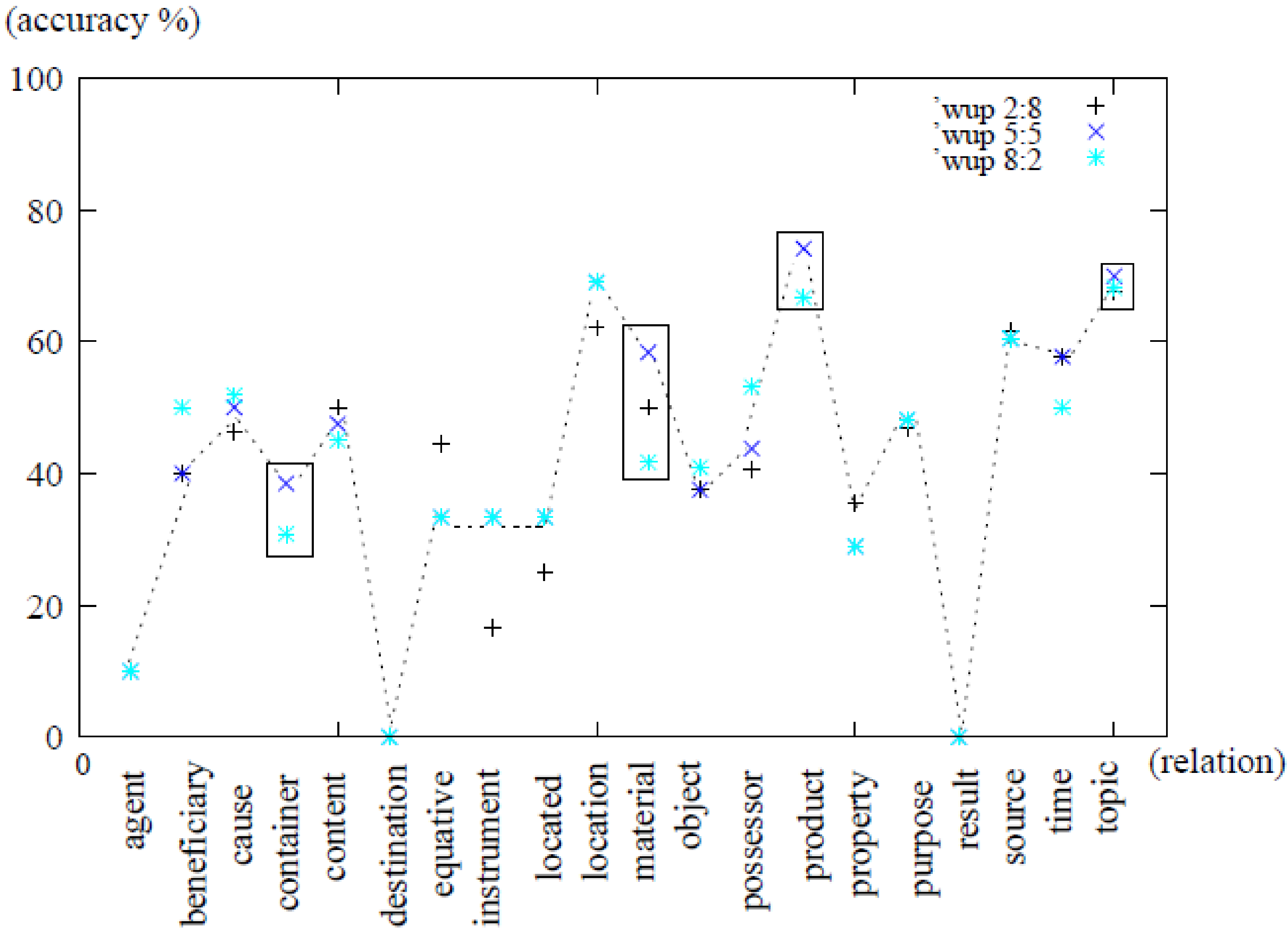


Fig. 4. Classifier accuracy at different α values

Experiment #2: Result

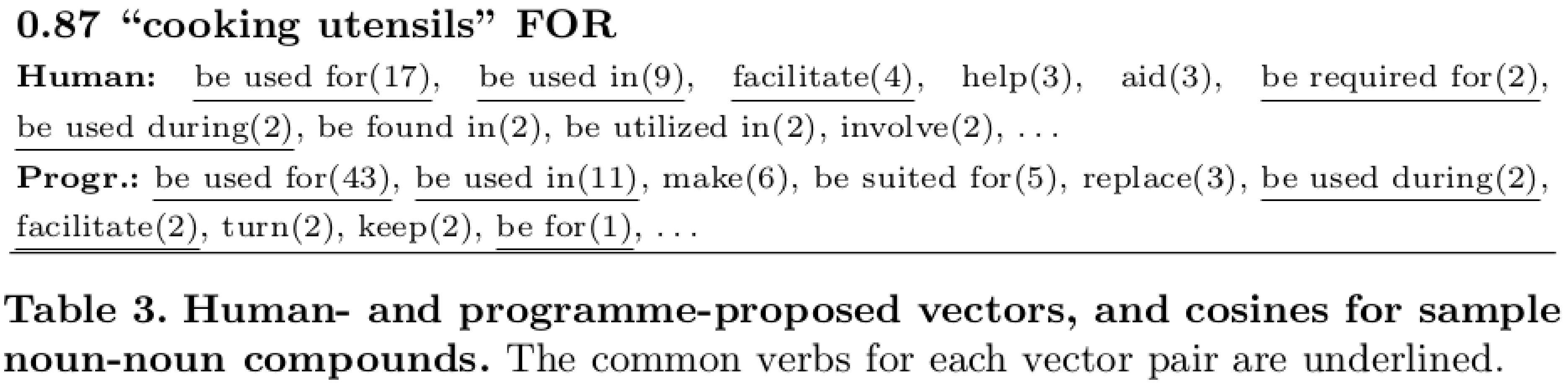


Various Relational Approaches

- Using 8 prepositions (Lauer, 1995)
- Verbs + prepositions (Nakov and Hearst, 2006)
- Using mind pattern from web (Turney, 2006)

  e.g. “*Y * couses X*” for CAUSE
- Pattern from corpus analysis (Turney & Littman, 2005)
  - 128 fixed phrases using 64 joining-terms

Relational Approaches: Example



0.87 "cooking utensils" FOR

Human: be used for(17), be used in(9), facilitate(4), help(3), aid(3), be required for(2), be used during(2), be found in(2), be utilized in(2), involve(2), . . .

Progr.: be used for(43), be used in(11), make(6), be suited for(5), replace(3), be used during(2), facilitate(2), turn(2), keep(2), be for(1), . . .

Table 3. Human- and programme-proposed vectors, and cosines for sample noun-noun compounds. The common verbs for each vector pair are underlined.

Use of Semantic Relation in NC

- Paraphrase-augmented machine translation
- Summarisation evaluation
- Textual entailment
- Information retrieval
  - index normalisation, query expansion, query refinement, results re-ranking, etc.
- Data mining
  - *Migraine treatment* → “ * *which prevents migraines*”

Our work

- Goal: extract “rules” for compound based on semantics of components
  - Used 20 relations porposed by Barker and Szpakowicz (1998)
- Explored ConceptNet, WordNet, and VerbNet
- Used CN2

References

- Judith N Levi. “*The syntax and semantics of complex nominals”*. Academic Press New York, 1978.
- Beatrice Warren. “*Semantic patterns of noun-noun compounds”. Acta Universitatis Gothoburgensis.* Gothenburg *Studies in English Goteborg*, 41:1–266, 1978
- Ken Barker and Stan Szpakowicz. “Semi-automatic recognition of noun modifier relationships”. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 96–102. Association for Computational Linguistics, 1998.
- Su Nam Kim and Timothy Baldwin. “A lexical semantic approach to interpreting and bracketing english noun compounds”. *Natural Language Engineering*, 19(03):385–407, 2013.
- Preslav Nakov. “On the interpretation of noun compounds: Syntax, semantics, and entailment”. *Natural Language Engineering*, 19(03):291–330, 2013
- Vivi Nastase and Stan Szpakowicz. “Exploring noun-modifier semantic relations”. In *Fifth international workshop on computational semantics (IWCS-5)*, pages 285–301, 2003

Thanks..