

Statistical Modeling of Multiword Expressions

Su Nam Kim

WING, CSSE Dept.



ChimeText seminar

Research Outline

Linguistics in MWEs

Statistical Approaches

Resources

Summary of Modeling Tasks

Summary

Research Outline

- **Aim:** to model the syntax and semantics of Multiword Expressions (MWEs) using statistical approaches
- **Significance**
 - ★ resolving the syntax and semantics of words as processing units
 - ★ number of MWEs is equivalent to simplex words (Jackendoff 1997)
 - ★ reusability (e.g. *take away/off/up..*), economics (e.g. *winter school*), new lexemes (e.g. *shock and awe, cell phone*), reliability and better expression (e.g. *piss me off*)
 - ★ fluency, robustness & better language understanding for NLP

Examples of English MWEs

1. (**NC**) The subject is about *language learning system design*.
2. (**VPC**) Kim *took* her pen *out*.
3. (**LVC**) She *took a* long *bath* for relaxation after taking a long exam.
4. (**Idiom**) He will inherit when his grandfather *kicks the bucket*.
5. (**D-PP**) The survey shows that *by and large* people skip breakfast.

Open Issues, Related Work & Limits

● Identification

- ★ determine whether multiple simplex words form a MWE in a given token context (*put the sweater on* vs. *put the sweater on the table*)
- ★ confusability with compositional expressions (e.g. *VPCs*, *LVCs*, *idioms*)

● Extraction

- ★ recognize MWEs as word units at type level
- ★ feed lexicon development

- **Detecting/Measuring Compositionality**

- ★ denote the degree of transparency among the components of MWEs
- ★ (assumption) meanings of MWEs and their parts are specified
- ★ hard to measure the degree of compositionality & to utilize it

- **Semantic Classification**

- ★ predict the semantics of MWEs involves understanding the degree of compositionality in MWEs
- ★ (assumption) meanings of MWEs are unspecified

- **Semantic Interpretation**

- ★ interpret the semantic association among components in MWEs
- ★ e.g. interpret the semantic relations in NCs, semantic classes of D-PPs such as media and manner
- ★ in case of NC interpretation, no standard set of SRs, conducted under artificial assumptions

- **Cross-over/Cross-lingual Study**

- ★ Utilize study outcomes of a language/type of MWE to another language/type of MWE
- ★ few cases of cross-lingual research; hard to find the same features among various MWE types (Venkatapathy 2006, Kim&Baldwin 2007)

Difficulties in Modeling MWEs

- syntactic, semantic, and pragmatic idiomaticity
 - ★ *family cars, He took off his coat, He kicked the bucket*
- syntactic and semantic flexibility
 - ★ *Eat dinner right up, make a big mistake*
- high productivity in language processing
 - ★ *orange/apple/lemon/chocolate... juice*
- different linguistic features w.r.t. various types of MWEs

Scope & Approaches of Thesis

● Scope of Research

- ★ English MWEs only
 - * due to resource availability
 - * focus on syntactically & semantically highly productive MWEs
- ★ Noun Compounds & Verb-Particle Constructions, due to the productivity and frequency of use

● Our Approach

- ★ use **Statistical** methods + symbolic methods
- ★ minimize human labor & maximize benefits of existing resources (e.g. WordNet, CoreLex)

Our Aim & Contribution

- to shed light on underlying linguistic processes giving rise to MWEs across constructions and across languages
- to generalize techniques, abstract away from individual MWE types to develop general purpose interpretation methods
- to cross-compare alignment of pre-existing MWE classifications
- exemplify the utility of MWE interpretation within general NLP tasks
- w/ **NCs**: NC interpretation, bracketing, WSD in NCs
- w/ **VPCs**: identification, detecting compositionality

Research Outline

Linguistics in MWEs

Statistical Approaches

Resources

Summary of Modeling Tasks

Summary

English MWEs: properties & types

- ***MWEs***: lexical items that can be decomposed into multiple simplex words and display lexical, syntactic, semantic, pragmatical and/or statistical idiosyncrasies
- **collocation and anti-collocation**
 - ★ collocation: any statistically significant word co-occurrence (Sag et al. 2002) (e.g. *red tape*)
 - ★ anti-collocation: a word which must *not* be used with the target words (Pearce 2001) (e.g. *many thanks* vs. *several thanks*)

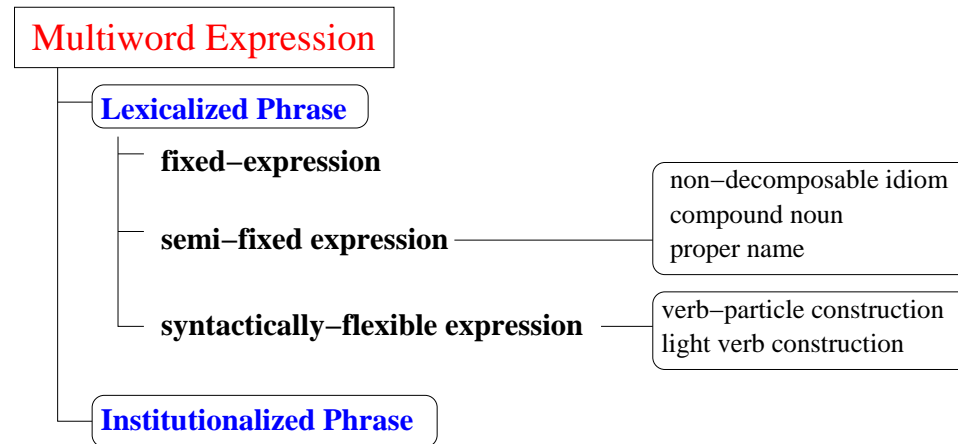
- **Properties of English MWEs** (Fillmore 1988, Liberman 1992, Nunberg et al. 1994, Sag et al 2002)
- ★ **Idiomatcity**: the syntactic, semantic, pragmatic, and statistical irregularity
 - * **syntactic idiomatcity**: e.g. *apple pie* vs ~~*by and large*~~
 - * **semantic idiomatcity**: e.g. *kick the bucket* vs. *bus driver*, related to non-identifiability, figuration
 - **Non-identifiability**: the meaning cannot be easily predicted from the surface form (components) (e.g. *kick the bucket* → *die??*)
 - **Figuration**: an attribute found in encoded expressions such as metaphors, metonymies and hyperboles (e.g. *red tape* = *bureaucratic*)

- * **pragmatic idiomaticity**: e.g. *good morning, all aboard*, related to situatedness
- * **statistical idiomaticity**: e.g. *black and white* vs. *white and black*, related to collocation
- ★ **Collocation** syntactically and semantically predictable but used with a high frequency in a particular context (e.g. *black and white* vs ~~*white and black*~~)

★ others

- * **Single-word paraphrasability**: paraphrasable MWEs enable substitution with a single word (e.g. *leave out* = *omit*)
- * **Proverbiality**: describe and implicitly explain a recurrent situation of particular social interest (e.g. *informality*, *affect*)
- * **Prosody**: have distinctive stress patterns that diverge from the norm (e.g. *soft spot* vs. *first aid* vs. *dental operation*)

- **Types of English MWEs** (Sag et al 2002)



- ★ fixed expression: no variation or internal modification
- ★ semi-fixed expression: lexically variable, various inflection, various reflexive form
- ★ institutionalized: statistically MWEs (e.g. *salt and pepper, many thanks, telephone booth, traffic light*)

Types of English MWEs

● Compound Nouns (CNs)

- ★ CN is a noun made up of two or more lexemes (cf. **NCs** = lexemes all nouns)
- ★ types: noun+noun(*morning tea*), adjective+noun(*monthly ticket*), noun+preposition(*hanger on*) etc.
- ★ Syntactic Variation: (*full moon*) vs. (*bed·room*) vs. (*post·man*, *post man*) vs. (*check-in*)
- ★ Modification such as plurality & genitive (*family cars* vs. *families car*)

● Verb-Particle Constructions (VPCs)

- ★ VPC is a verb with obligatory particle(s)
 - ★ **intransitive:** *Kim calmed down.*
 - ★ **transitive:** *Kim handed in the paper./Kim handed the paper in./Kim gets Sandy down.*
- **Light-Verb Constructions (LVCs)**
 - ★ LVC is a verb whose meaning is bleached to some degree & whose noun complement dominates semantic determination
 - ★ Occur in many languages such as English, Dutch, Japanese
 - ★ Common English light verbs: *do, get, give, have, make, put, take*
 - ★ Examples (Butt 2003): *do a memo, take a bath, give a bath, make a decision*

- **Idioms**

- ★ MWE whose meaning is not predictable from the usual meanings of its parts
- ★ categorized into *compositional* (*take advantage of, spill the beans*) vs. *non-compositional* (*in one's shoes, kick the buckets*)

- **Determinerless-Prepositional Phrase (D-PPs)**

- ★ MWE constructed with a preposition & a singular noun w/o a determiner
- ★ Syntactic Markedness: (non-)productive (e.g. *by car/bus..* vs. *at large/small/medium*) & (non-)modifiable (e.g. *per person* w/ countable NN vs. *at school* w/ uncountable NN)
- ★ Nominal Modifiability : fully fixed expressions (*of very course*)

vs. obligatory modification(*at great expense*)

★ Semantically Markedness

semantics	examples
institutional	<i>at school, in church, on campus, in gaol</i>
media	<i>on TV, on record, off screen, in radio</i>
metaphor	<i>on ice, at large, at hand, at liberty</i>
temporal	<i>at breakfast, on holiday, on break, by day</i>
means/manner	<i>by car, by hammer, by computer, via radio</i>

Research Outline

Linguistics in MWEs

Statistical Approaches

Resources

Summary of Modeling Tasks

Summary

Statistical Approaches

● Co-occurrence Properties

- ★ Use the co-occurrence of parts in the target MWE for task
- ★ Work with **collocation** & **anti-collocation** implicitly
- ★ Example: *propose* & co-occurring words (Lin 1999)
 - * million, billion, accord, increase, call, year, change, support, proposal, percent, money, plan, cut, aid, program, people

● Substitutability

- ★ The ability to replace parts of lexical items with alternatives
- ★ when parts in lexical items occur w/ unusually high frequency
- ★ Work with **collocation** & **anti-collocation** explicitly

- ★ MWEs and Non-MWEs using substitutability (Pearce 2001)

- * *frying fan* → *frying **pot***, *salt and pepper* → *salt and **sugar***, *many thanks* → ***some** thanks*

- **Distributional Similarity**

- ★ When two words are similar, their context words are also similar

- ★ Example, *kick the bucket*:

- * (MWE:mourn, sad, blue) vs. (Non-MWE:run, ball, accident)

- **Semantic Similarity**

- ★ Underlying assumption of semantic similarity: the similarity of the parts represents the semantics of the whole

- ★ Examples w/ NCs : modifier = *fruit*, head noun = *liquid* (SR:MAKE) e.g. *apple juice, orange juice, grape nectar*

● Ellipsed Predicates

- ★ a way to use the semantic association of parts while building constructs which put parts together
- ★ correlated with compositionality
- ★ w/ NC, *virus infection* → SR, CAUSE (Levi 1979)
 1. infection (virus causes infection)
 2. infection (infection is caused by virus) → Passive
 3. infection (infection is virus-caused) → Compound adjective Formation
 4. ...

● Linguistic Properties

- ★ Linguistic features can be strong clues for lexical acquisition
- ★ Examples w/ VPCs : **particle position** *pick it up.* vs. ~~*pick up it.*~~

Research Outline

Linguistics in MWEs

Statistical Approaches

Resources

Summary of Modeling Tasks

Summary

Resources (1)

- Corpus
 - ★ British National Corpus
 - ★ Brown Corpus
 - ★ Wall Street Journal in Penn Treebank

Resources (2)

● Lexical Resources

★ WordNet

- * lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory (Fellbaum:1998)

★ Moby's Thesaurus

- * based on Roget's Thesaurus, contains 30K root words and 2.5M synonyms and related words

★ CoreLex

- * systematic polysemy and semantic underspecification of nouns from WordNet 1.5 (Buitelaar 1998)

Resources (3)

- Tools

- ★ WordNet::Similarity

- * **Relatedness** has-part, is-made-of, is-an-attribute-of (lesk, vector)

- * **Similarity:path-based** is-a (wup, lch, path)

- * **Similarity:information-based** is-a (jcn, lin, lesk)

- * **Random** (random)

- ★ TiMBL

- * machine learner (Daelemans 2004)

- ★ RASP parser

- * extract argument structure from the output of the dependency analysis

Research Outline

Linguistics in MWEs

Statistical Approaches

Resources

Summary of Modeling Tasks

Summary

Summary of Modeling Tasks

- Constituent similarity method using semantic similarity (IJCNLP05)
 - ★ similar NCs tend to have the same SR
 - ★ e.g. *apple juice*:MAKE → *banana milk*:MAKE
- Verb semantics method based on underlying predicate (ACL06)
 - ★ Using the verb semantics defined in Semantic Relations and grammatical role of head noun and modifies
 - ★ e.g. *GM car*=MAKE → *car made by GM*

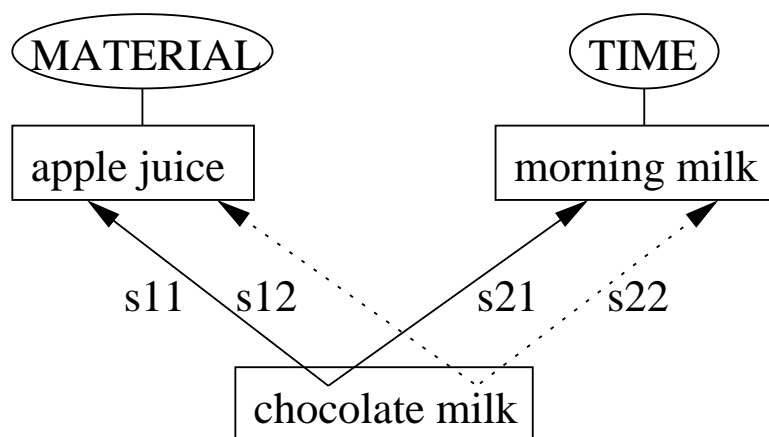
- Constituent substitution method using **substitutability, semantic similarity, co-occurrence** (PACLING07)
 - ★ expand the interpreted NCs by substitution based on sense collocation and bootstrapping
 - ★ e.g. *apple juice:MAKE* → *fruit/crabapple/orange juice:MAKE*
- Benchmarking & hybridizing NC interpretation methods using **semantic similarity, substitutability, co-occurrence** (SemEvalworkshop07, IJCNLP08)
 - ★ with sense collocation, constituent similarity and constituent substitution methods, hybridize and benchmark using SEMEVAL-2007 data

- WSD in NCs using **substitutability, semantic similarity** (AAAI07)
 - ★ use sense collocation, roles of parts and heuristics (**one sense per collocation**)
 - ★ e.g. $(\text{TOPIC} | WS_{art}, WS_{museum}) \rightarrow (WS_{art} | WS_{museum}, \text{TOPIC} / \text{grammatical_role}_{art})$
 - ★ e.g. *art museum* \rightarrow artifact/creation/skill/visual museum
- Identify VPCs using **linguistic properties** (EACLworkshop06)
 - ★ use linguistic properties of associated nouns of VPCs and Verb-PPs associated with distinct **selectional preferences**
 - ★ e.g. put the coat on vs. put the coat on the chair

- Detect compositionality of VPCs using **semantic similarity** (PACLING07)
 - ★ use semantic similarity of combination of Verb and Particle
 - ★ e.g. *call up* = compositional → *ring up* = compositional

Using Constituent Similarity

- **Intuition:** Similar NCs have the same SR



	Training	Test	S_{ij}
n_1	apple	chocolate	0.71
n_2	juice	milk	0.83
n_1	morning	chocolate	0.27
n_2	milk	milk	1.00

Figure 1: *w/ chocolate milk*

Method

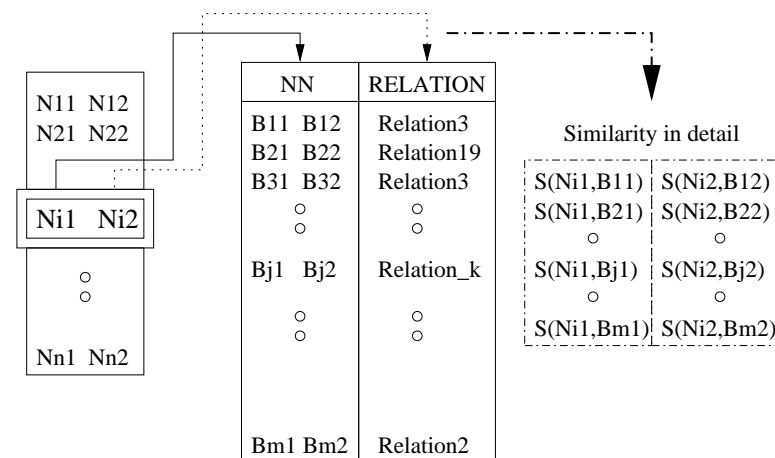
- Compute the Similarity

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

- Find the SR for the test NC

$$\star rel(N_{i,1}, N_{i,2}) = rel(B_{m,1}, B_{m,2}) \text{ where } m = \operatorname{argmax}_j S((N_{i,1}, N_{i,2}), (B_{j,1}, B_{j,2}))$$

- Similarity between i_{th} NC in test NC and j_{th} NC in training NC



Summary of Constituent Similarity Method

- Achieved higher performance than previous results
- Confirm the relative contribution of parts w.r.t. SRs
- test the method over 3-term NCs
- Successfully adopt other techniques (bootstrapping & k -NN)
- Show the utility of SRs in bracketing
 - ★ e.g. *((computer science) department)* vs. *(linguistic (graduate proram))*

NC interpretation via Verb Semantics (1)

- Using the verb semantics defined in Semantic Relations and grammatical role of head noun and modifier

(1) *family car*

case: family owns the car.

form: H own M

relation: POSSESSOR

(2) *student protest*

case: protest is **performed** by student.

form: M is performed by H

relation: AGENT

(3) *family car*

case: *Synonym=have/possess/belong to*

form: H own M

relation: POSSESSOR

(4) *student protest*

case: *Synonym=act/execute/do*

form: M is performed by H

relation: AGENT

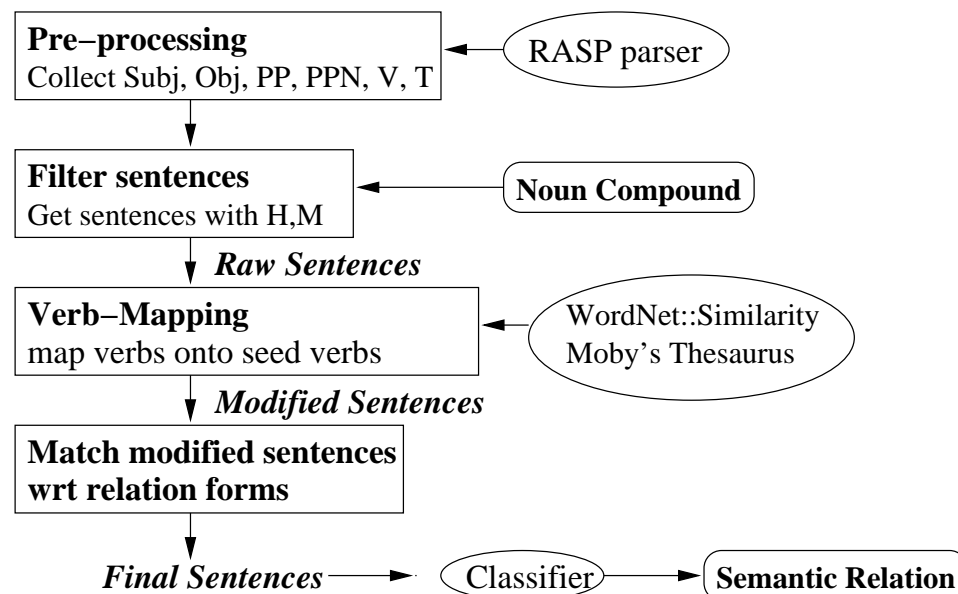
NC interpretation via Verb Semantics (2)

- **Emerging issue:** can we have enough instances for interpretation?
- **Solution:** map actual verbs onto verb classes in terms of SRs based on **seed verbs**
- *What are **seed verbs**?*
 - ★ verbs from definition of SRs and some of their synonyms
 - ★ two sets of seed verbs (57 vs. 84)
 - ★ example of seed verbs for SR POSSESSOR:
 - (57) own/have/possess/belong to
 - (84) own/have/possess/belong to/acquire/grab/occupy

Method & Architecture

- Example of constructional templates associated with SR, POSSESSOR

★ $S(\text{have, own, possess}_{verb}, M_{subj}, H_{obj}), S(\text{belong_to}_{verb}, H_{subj}, M_{obj})$



Summary of Verb Semantics Method

- Achieved 52.63% with 84 seed verbs using **VECTOR** mapping method from **Weight**
- Investigate the effective verb mapping method to expand the instances
- Test two different sets of seed verbs
- Outperformed Moldovan (2004) and Kim & Baldwin (2005)
- Show performance of similarity method introduced by Kim & Baldwin (2005) over our data set

Word Sense Disambiguation for NCs

- **Aim:** to investigate the interaction between word sense and interpretation in English NCs
 - ★ to automatically disambiguate polysemous nouns in NCs
 - ★ to improve NC interpretation performance through word sense

Sense Distribution

- The sense distribution of nouns in NCs differs from simplex usages
- The sense distribution of modifier and head nouns also differs, e.g. *art* and *day* (based on SemCor and WordNet2.1):

WordNet sense	<i>art</i>		
	mod	head	SemCor
WS ₁	.85	.62	.67
WS ₂	.11	.04	.22
WS ₃	.00	.03	.08
WS ₄	.04	.31	.03

WordNet sense	<i>day</i>		
	mod	head	SemCor
WS ₁	.13	.04	.41
WS ₂	.02	.04	.20
WS ₃	.80	.00	.12
WS ₄	.00	.91	.20
WS ₅	.04	.01	.05
WS ₆	.00	.00	.03

One Sense per Collocation

- One Sense per Collocation heuristic of Yarowsky (1995)
 - ★ words almost always occur with the same sense across all token instances of a given word collocation
 - ★ accuracy of 90-99% over a range of binary disambiguation bootstrapping tasks
- One Sense per Collocation for NC
 - ★ apply the heuristic to the full WordNet sense inventory rather than coarse-grained binary distinctions
 - ★ apply to NCs at the type level (i.e. no linguistic claims made for different senses based on context)

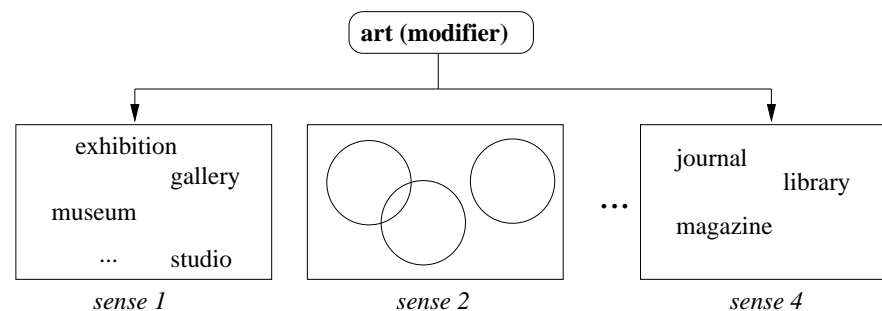
Approach I: Supervised (1)

- Use sense combinatoric method of Moldovan et al. (2004):

$$sr^* = \operatorname{argmax}_{sr_i} P(sr_i | ws(n_1), ws(n_2)) \quad (1)$$

$$ws^*(n_i) = \operatorname{argmax}_{ws(n_i)} P(ws(n_i) | ws(n_j), sr/gr) \quad (2)$$

- Replace sr in (2) with the **grammatical role** of the polysemous noun (gr) (e.g. art museum vs. martial art)



Approach I: Supervised (2)

- Experiment with two sense inventories:
 - ★ CoreLex, e.g. *apple* = FOOD (61.6% coverage)
 - ★ first sense and its three hypernyms in WordNet2.1

art	lesson
Sense 1	lesson => teaching, instruction,... => education => profession => ...
Sense 2	example, deterrent example..

Approach II: Unsupervised

- Replace a polysemous noun with its synonyms and calculate the probability of each underlying word sense by web frequency:

$$ws^*(n_1) = \operatorname{argmax}_{s_i \in ws(n_1)} \frac{\sum_{n_j \in ss(s_i) \setminus \{s_i\}} \operatorname{freq}(n_j, n_2)}{|ss(s_i) \setminus \{s_i\}|}$$

- Example of substitution method with art museum

sense	substituted NCs
1	craft/artifact museum
2	artistic production/creative activity museum
3	artistry/superior skill museum
4	artwork/graphics/visual communication museum

Summary of WSD in NCs

- The proposed (supervised) WSD method works well over NCs
 - ★ best performance = 55% accuracy
 - ★ tested semantics of non-polysemous nouns → first sense and its hypernyms is more practical choice
- Off-the-shelf WSD methods do not apply well to MWEs
 - ★ SENSELEARNER performed poorly over NCs (accuracy = 30%)
- WSD improves NC interpretation performance
 - ★ indication there is room for more improvement

Research Outline

Linguistics in MWEs

Statistical Approaches

Resources

Summary of Modeling Tasks

Summary

Summary (reading list)

- related to **NC interpretation**

1. **Su Nam Kim**, Timothy Baldwin, *Automatic Interpretation of Semantic Relations in Compound Nouns using WordNet Similarity*, 2nd International Joint Conference on Natural Language Processing (IJCNLP), 2005, Jeju island, Republic of Korea, pp.945–956
2. **Su Nam Kim**, Timothy Baldwin, *Interpreting Semantic Relations in Noun Compounds via Verb Semantics*, The Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (Coling/ACL), 2006, Sydney, Australia, pp.491–498
3. **Su Nam Kim**, Timothy Baldwin, *MELB-KB: Nominal Classification as Noun Compound Interpretation*, 4th International Workshop on Semantic Evaluations (SemEval), 2007, Prague, Czech Republic, pp.231–236
4. **Su Nam Kim**, Timothy Baldwin, *Interpreting Noun Compound using Bootstrapping and Sense Collocation*, Conference of the Pacific Association for Computational Linguistics (PACLING), 2007, Melbourne, Australia, 129-136
5. **Su Nam Kim**, Timothy Baldwin, *Benchmarking Noun Compound Interpretation*, 3rd International Joint Conference on Natural Language Processing (IJCNLP), 2008, Hyderabad, India
6. **Su Nam Kim**, Meladel Mistica, Timothy Baldwin, *Australian Language Technology Workshop*, Melbourne, Australia
7. **Su Nam Kim**, Timothy Baldwin, *Noun Compound Interpretation : Feasibility Study of Syntax and Semantics in Noun Compounds*, Journal of Natural Language Engineering (NLE), Cambridge (in preparation)

- related to **VPC**

1. **Su Nam Kim**, Timothy Baldwin, *Automatic Extraction of Verb-Particles Using Linguistic Features*, 11th Conference of European Chapter of the Association for Computational Linguistics : 3rd ACL-SIGSEM Workshop on Preposition, 2006, Trento, Italy, pp.65–72
2. **Su Nam Kim**, Timothy Baldwin, *Detecting Compositionality of English Verb-Particle Constructions using Semantic Similarity*, Conference of the Pacific Association for Computational Linguistics (PACLING), 2007, Melbourne, Australia, pp.40-48
3. **Su Nam Kim**, Timothy Baldwin, *Identifying English Verb-Particle Constructions via Linguistic Features*, Special issue of the International Journal of Language Resources and Evaluation (LRE) (under review)

- related to **WSD**

1. **Su Nam Kim**, Timothy Baldwin, *Disambiguating Noun Compound*, 22nd AAI Conference on Artificial Intelligence (AAAI), 2007, British Columbia, Canada, pp.901-906
2. David Martinez, **Su Nam Kim**, Timothy Baldwin, *MELB-MKB:Lexical Substitution system based on Relatives in Context*, 4th International Workshop on Semantic Evaluations (SemEval), 2007, Prague, Czech Republic, pp.237–240
3. Timothy Baldwin, **Su Nam Kim**, Francis Bond, Sanae Fujita, David Martinez and Takaaki Tanaka, *MRD-based Word Sense Disambiguation: Further Extending Lesk*, 3rd International Joint Conference on Natural Language Processing (IJCNLP), 2008, Hyderabad, India
4. Timothy Baldwin, **Su Nam Kim**, Francis Bond, Sanae Fujita, David Martinez and Takaaki Tanaka, *A Reexamination of MRD-based Word Sense Disambiguation*, ACM Transactions on Asian Language Information Processing (in preparation)

Direction of Future Research

- Expand the investigated methods for better performance
- Integrate outcomes into NLP applications & crossover/crosslingual study
- Related to NCs:
 - ★ investigate unsupervised methods
 - ★ determine & propose a reliable set of SRs along with comparison methods
 - ★ deal with SR pragmatism

- ★ utilize the research outcome into a real-world NLP applications
- Related to VPCs:
 - ★ investigate unsupervised methods to extract/identify VPCs
 - ★ deal with the measure of degree of VPC compositionality
 - ★ utilize the research outcome into a real-world NLP applications

Notes

- I will present two selected papers in near future below:
 - ★ **Su Nam Kim**, Timothy Baldwin, *Interpreting Semantic Relations in Noun Compounds via Verb Semantics*, The Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (Coling/ACL), 2006, Sydney, Australia, pp.491–498
 - ★ **Su Nam Kim**, Timothy Baldwin, *Disambiguating Noun Compound*, 22nd AAI Conference on Artificial Intelligence (AAAI), 2007, British Columbia, Canada, pp.901-906

Notes

Thank you for your attention.
Questions and/or Suggestions?!