공학적 관점에서의 온톨로지 구축
(Ontology Construction from the Aspect of Engineering)

Ko, Youngjoong *(yjko@dau.ac.kr)*

Dept. of Computer Engineering
Dong-A University

http://web.dong.ac.kr/yjko/

Contents

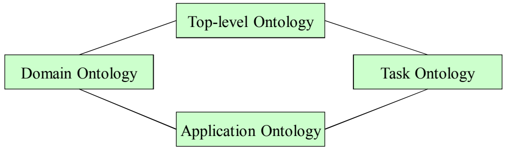
- Ontology and Information Systems
  - Engineering Ontologies
  - Linguistics Ontologies
  - Ontologies for Web Applications
  - Ontologies for NLP Applications
- Case Study: The Omega Ontology from ISI in USC
  - Methodologies for the Reliable Construction of Ontological Knowledge
  - Combining of Large-Scale, Practical Ontologies
  - Enriching Ontologies Using the WWW

Introduction

- Ontology
  - Confined to the philosophical sphere in the past
  - Hierarchically organized networks of conceptual information
  - Used to systematize and model domain knowledge, they play an important role in *Artificial Intelligence*, *Natural Language Processing*, *Information Integration*, *Electronic Commerce*, and related fields

Ontologies and Knowledge Bases

- A Classification of Ontologies (Guarino, 1997)



  - Top-level Ontology
    - Generic concepts independently of a particular domain or problem
    - Fairly abstract categories of time, space, event, action, etc
    - e.g. Mikrokosmos (Mahesh, 96), Sensus (Knight & Luk, 94)
  - Domain Ontology
    - the terminology of a generic domain
    - e.g. UMLS Metathesaurus (National Library of Medicine, 1997)

Ontologies and Knowledge Bases

- Task Ontology
  - Generic Tasks or Activities
  - e.g. Bank transactions, diagnosing
- Application Ontology
  - Combination or specialization of task and domain ontologies

- Knowledge Bases
  - Ontologies describe possible worlds.
  - A knowledge base describes an instance of a possible world as a set of facts.
  - Knowledge base repository records
    - Specific events, places, people, objects, etc., as classified by the ontology

Engineering and Linguistic Ontologies

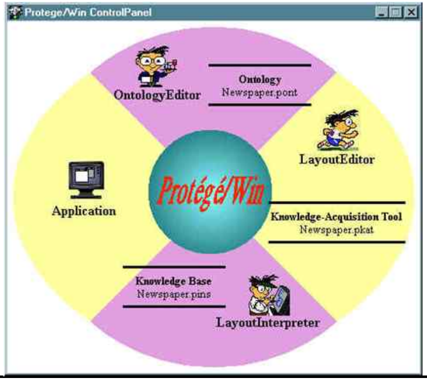
- Engineering Ontologies for Information Systems (IS)
  - Grounded in some view of the real world
    - e.g. a world of customers and banking transactions
  - Derive software that interact with the world according to this view
    - e.g. process a bank teller transaction
  - Application ontologies
    - Typically used at development time to constraint or help IS designers specify the schema of the application
    - Helps the analyst design the IS such as CASE tools
      - e.g. modeling a product and customer DB, selling transactions
    - Used to generate a software program from high-level formal specifications
      - e.g. automating customer requests, bank teller transactions, inventory management etc.

Engineering and Linguistic Ontologies

- A Example Tool for an Engineering Knowledge Base
  - ***Protégé/Win*** from Stanford Univ.
  - A suite of software tools used by system developers and domain experts to develop *knowledge-based systems*
    - One module to construct an *ontology* of abstract classes
    - Second module to create a *knowledge-acquisition tool* for collecting knowledge
    - Third module to enter specific instances of data and create a *knowledge base*
  - Reuse domain ontologies and problem-solving methods
    - Shortening the time needed for development and program maintenance

Engineering and Linguistic Ontologies

- A Newspaper Example



Linguistic Ontologies and NLP

- Language Ontologies
  - An ontology for Natural Language Processing (NLP)
    - A set of concepts, properties of concepts, relationships between concepts
      - Used for the purpose of building a *semantic representation* from an input text analysis or create a text from a *semantic expression generation*
    - Encode a common sense view of the world
    - Provide a representation of linguistic meaning using extra-linguistic terms and expressions
  - Disambiguated semantic representation
    - So-called selectional restrictions are used to select the appropriate sense of a word by checking the semantic type

Ontologies for the Web applications

- Ontologies for the Web or for Documentation Systems
  - Typically larger and a simpler structure than engineering or linguistic ontologies
  - Applications include:
    - Structured index with concept hierarchy and other links (part-of, function, etc.)
    - Site structuring for supporting browsing and localized search
    - Meta-tagging (e.g. topic tagging using a terminology) and query expansion
    - Data/web-mining

Ontologies for the Web applications

- The Role of Ontology for the Web
  - Annotation of text in a specialized Web crawler
  - Deriving answers in the search engine
  - Ontology-based search is used to solve a set of specific problems
    - Unfamiliar vocabulary
      - Provide a rich network of terms
    - Word Sense Disambiguation
    - Retrieval of short documents
      - Query expansion

Ontologies for the NLP applications

- Ontology Content: Shallow Semantics
  - "shallow semantics"
    - Building adequate ontologies at the time was impossible
    - Help statistical NLP systems overcome the quality performance ceilings many of them seem to have reached
      - Machine Translation, Text Summarization, Information Retrieval, Question Answering, Dialogue Management etc.
  - A semantic theory needed to support NLP and other applications
    - A collection of unambiguous semantic symbols
    - Each carries a clear denotation
    - A set of rules for composing these symbols
    - Some method of validating the results of composition, deduction, and other semantic operations

Ontologies for the NLP applications

- What's ontologies in NLP engineering?
  - Sets of symbols taxonomized to enable inheritance of information and to support inference
    - e.g. WordNet
  - To denote any set of terms organized hierarchically according to the general property inheritance relation following subclass
  - Terminology taxonomies such as WordNet
    - Support knowledge representation needs in some practical application
    - More concerned about the *computational effectiveness and correctness* of their application than about the formal completeness, correctness, or consistency of the ontology

Ontologies for the NLP applications

- In NLP applications
  - Build ontologies as relatively simple term taxonomies with some inheritance inference
  - Not enforce stricter logical requirements

Case Study: The Omega Ontology

Ko, Youngjoong

Dept. of Computer Engineering
Dong-A University

The Methodology of Ontology Construction

- Ontology Construction
  - Start with the ontologies of others and combine, prune, and massage them together as needed
  - Still lacking in ontology construction
    - Systematic and theoretically motivated methodology
      - To guides the builder and facilitates consistency and accuracy at all levels
      - Not have and adequate theory on which to base such a methodology
    - e.g. The OntoSelect Website (http://views.dfki.de/ontologies/)
      - Not one of the builders would be able to provide a set of operationalizable test

The Methodology of Ontology Construction

- Five Types of Research Approaches to Build Ontolgogies
  - Five types of motivation to construct ontologies
    - *Philosophers*
    - *Cognitive Scientists*
    - *Linguists*
    - *Artificial Intelligence Reasoners including Computational Linguistics*
    - *Domain Specialists*

    - Operates in a distinct way, resolving questions with arguments that appeal to different authorities and patterns of reasoning, and lead to very different results

The Methodology of Ontology Construction

- To Generate New Ontology
  - Ontologizer
    - *Whether to create a term*
    - *How to place it with regard to the other existing terms*
    - *Additional specification and definition*
  - This decision process plays out for five 'personality types' of ontologizer

- **Type 1**: Abstract feature recombination (the philosophers)
  - The historical method of ontologies
  - Modern Version
    - Define several highly abstract features
    - More or less mechanically forms combinations of abstract features as concepts, using these features as differentiae
    - DOLCE ontology (http://www.loa-cnr.it/DOLCE.html)
  - Elegant, but unfortunately doesn't work beyond the very most abstract level
    - Not very useful for practical domain ontologies

The Methodology of Ontology Construction

- **Type 2**: Intuitive Ontologies Distinctions (the cognitive scientists)
  - Methodology to determine concept formation
    - Devise clever experiments to measure how people make distinctions between close concepts
  - The fluidity of the distinction process
    - Dependant on the person's interests, knowledge, task, and other circumstances
    - Make this approach to ontology building fraught with inconsistence to the point of hopelessness

The Methodology of Ontology Construction

- **Type 3**: Cross-linguistic Phenomena (the linguists)
  - Some NLP applications
    - Pay attention to many languages for ontology construction and lexicon development can be rewarding
    - Many cultures independently name a thought
    - e.g. EuroWordNet
  - No one will accept the argument that "because it's so in languages, it has to be exactly so in thought"

The Methodology of Ontology Construction

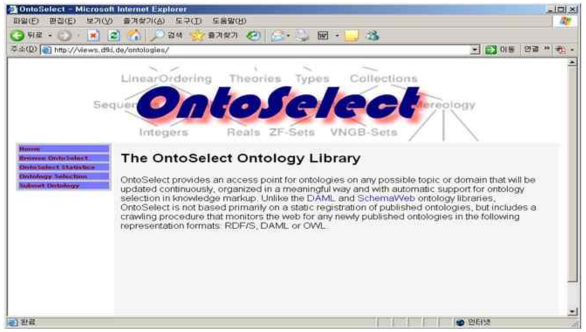
- **Type 4**: Inference-based Concept Generalizations (the Computational Reasoners)
  - The domain terms
    - The terms in ontology of some domain are grouped together.
    - Such groupings tend to emphasize domain-specific concepts and produce more abstract concepts only as they are required for grouping.
    - Tend to mirror the metadata and the system variables
  - Relatively clean, depending on the elegance of the computational solution to the problem
  - But, decision justifications are seldom interesting to philosophers, psychologists, and linguistics

The Methodology of Ontology Construction

- **Type 5**: Inherited domain distinctions (the domain specialists)
  - In many ontology building enterprises
    - The reason for creating and arranging concepts
      - Not from abstract theoretical analysis or experimentation, but from existing domain theory and practice
    - Biologists, neuroscientists, aircraft builders, pump manufacturers, legal scholars, and anyone in knowledge-intensive enterprises
      - Find it perfectly natural to construct ontologies that reflect the way their fields view their worlds
      - e.g. The OntoSelect Website (http://views.dfki.de/ontologies/)

The Methodology of Ontology Construction

- The OntoSelect Website


OntoSelect - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
뒤로 검색 즐겨찾기
주소(D) http://views.dfki.de/ontologies/
이동 연결
LinearOrdering Theories Types Collections
Sequen OntoSelect Mereology
Integers Reals ZF-Sets VNGB-Sets
Home
Browse OntoSelect
OntoSelect Statistics
Ontology Selection
Submit Ontology
완료
인터넷

The OntoSelect Ontology Library

OntoSelect provides an access point for ontologies on any possible topic or domain that will be updated continuously, organized in a meaningful way and with automatic support for ontology selection in knowledge markup. Unlike the DAML and SchemaWeb ontology libraries, OntoSelect is not based primarily on a static registration of published ontologies, but includes a crawling procedure that monitors the web for any newly published ontologies in the following representation formats: RDF/S, DAML or OWL.

Ontology Construction Procedure

- Continual Graduated Refinement

1. Determine the general characteristics of the ontology to be built
   - The domain of interest, the purpose of the ontology, the target level of granularity, the conceptual and theoretical antecedents
2. Gather all additional knowledge resources
   - Starter ontologies, upper structure or micro-theories, glossaries of domain terms, supporting descriptive and definitional material, algorithms and tools, existing theoretical descriptions
3. Delimit the major phenomena for consideration
   - Identify the core concepts, types of features allowed, principal differentiae
   - *Starting with an existing upper ontology (helpful)*

Ontology Construction Procedure

4. List all readily apparent terms/concepts important for the task or enterprise
   - Derived from a data model, from the algorithm of the system built, and from experts' reports on the major components and processes in the domain

5. For each concept, explicitly record the principle and factors that justify its creation
   - *Although the definition may still be incomplete and informal, but should contain the principal differentiae and features of interest*
   - *Identify interrelationships between the concept and related concepts*

6. Inspect the nascent domain model for regularity, balance, etc
   - For each major region repeat steps 3 to 5. refining existing concepts
   - During this iterative refinement, record all problematic issues

7. When done, characterize the ontology or domain model by recording its essential parameters

Combining Large-scale, Practical Ontologies

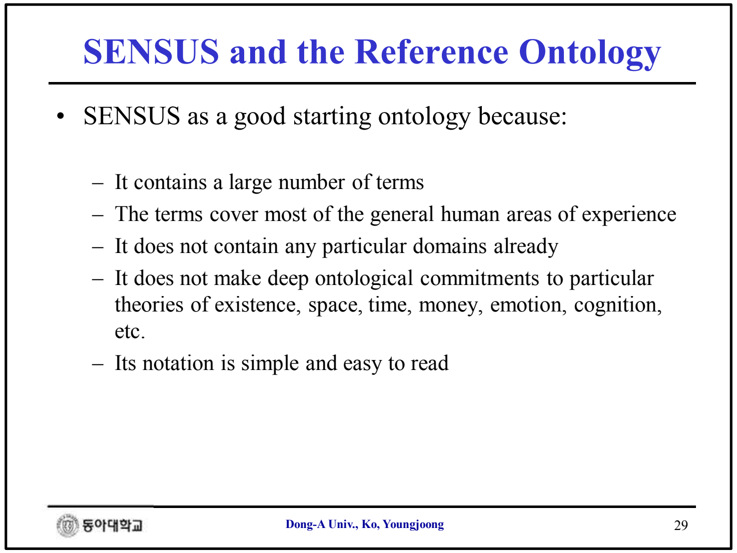
- The need of some neutral internal terminology
  - Incommensurate views of the 'same' object mean incompatible systems
  - The only solution is to try to map terms to each other
    - Directly mapping using large bi-domain correspondence tables
      - N^2 mapping sets
    - Indirectly mapping using some neutral internal terminology
      - *N* mapping sets
  - The advantages of a single, neutral ontology
    - They help standardize terminology
    - They assist knowledge transfer
    - They facilitate interoperability

SENSUS and the Reference Ontology

- Creating a standard ontology
  - Begin with a large, high-level, but rather content-neutral ontology
  - Systematically add to it other ontologies, term-sets, data definitions

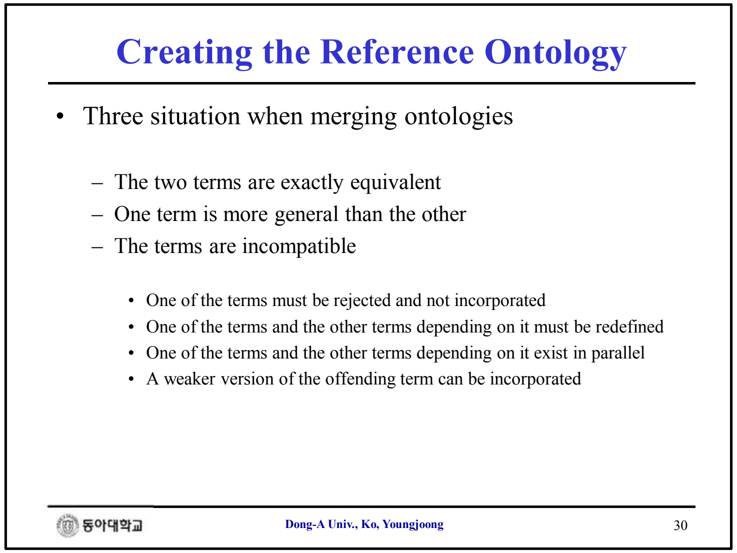
SENSUS and the Reference Ontology

- Initial central and high-level ontology
  - SENSUS built at USC/ISI
    - Serve as the internal mapping structure between lexicons of Japanese, Arabic, Spanish, and English
      - GAZELLE: machine translation
      - SUMMARIST: multilingual text summarization
      - C*ST*RD: multilingual text retrieval and management
    - Contains approx. 70,000 terms
      - Linked together into a subsumption (isa) network (including part-of, pertains-to etc.)
      - Japanese (120,000 root words), Arabic (60,000), Spanish (40,000), English(90,000)



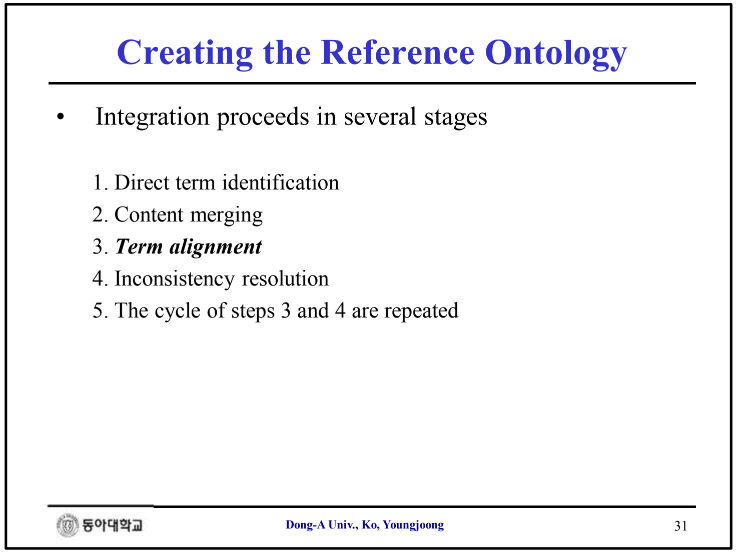
SENSUS and the Reference Ontology

- SENSUS as a good starting ontology because:
  - It contains a large number of terms
  - The terms cover most of the general human areas of experience
  - It does not contain any particular domains already
  - It does not make deep ontological commitments to particular theories of existence, space, time, money, emotion, cognition, etc.
  - Its notation is simple and easy to read



Creating the Reference Ontology

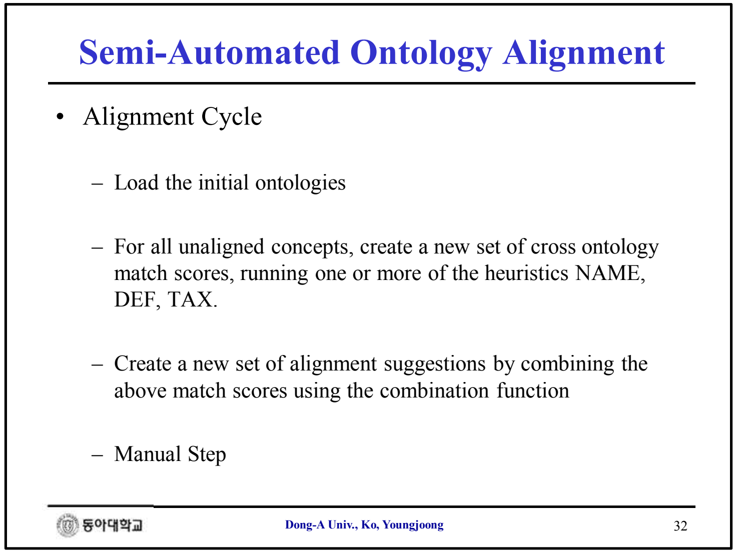
- Three situation when merging ontologies
  - The two terms are exactly equivalent
  - One term is more general than the other
  - The terms are incompatible
    - One of the terms must be rejected and not incorporated
    - One of the terms and the other terms depending on it must be redefined
    - One of the terms and the other terms depending on it exist in parallel
    - A weaker version of the offending term can be incorporated



Creating the Reference Ontology

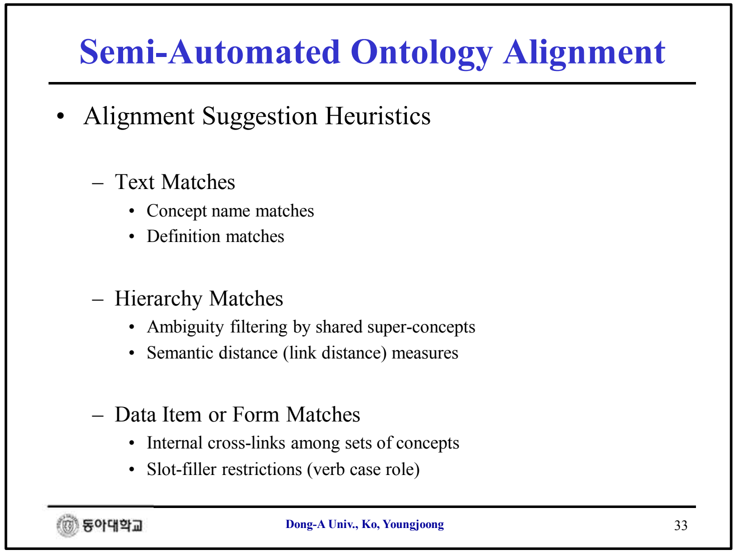
- Integration proceeds in several stages

1. Direct term identification
2. Content merging
3. ***Term alignment***
4. Inconsistency resolution
5. The cycle of steps 3 and 4 are repeated



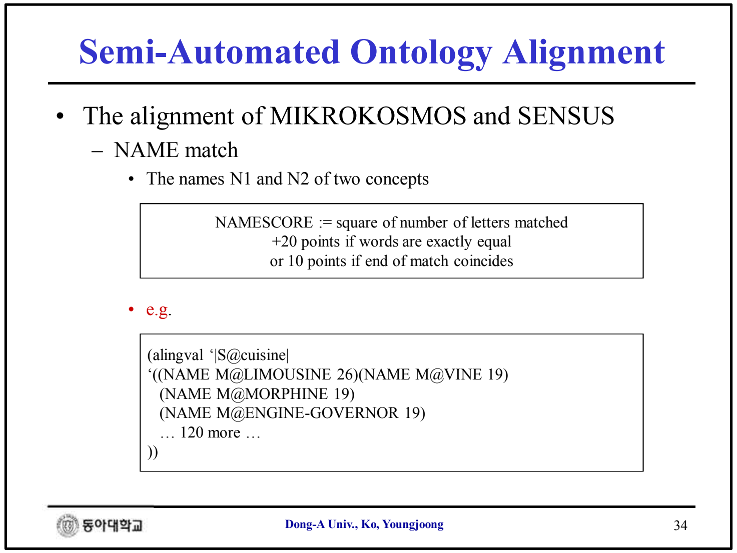
Semi-Automated Ontology Alignment

- Alignment Cycle
  - Load the initial ontologies
  - For all unaligned concepts, create a new set of cross ontology match scores, running one or more of the heuristics NAME, DEF, TAX.
  - Create a new set of alignment suggestions by combining the above match scores using the combination function
  - Manual Step



Semi-Automated Ontology Alignment

- Alignment Suggestion Heuristics
  - Text Matches
    - Concept name matches
    - Definition matches
  - Hierarchy Matches
    - Ambiguity filtering by shared super-concepts
    - Semantic distance (link distance) measures
  - Data Item or Form Matches
    - Internal cross-links among sets of concepts
    - Slot-filler restrictions (verb case role)



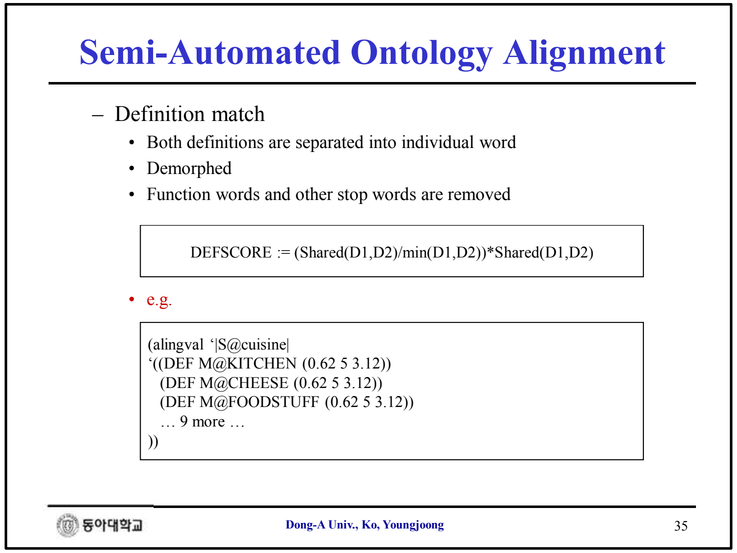
Semi-Automated Ontology Alignment

- The alignment of MIKROKOSMOS and SENSUS
  - NAME match
    - The names N1 and N2 of two concepts

NAMESCORE := square of number of letters matched
+20 points if words are exactly equal
or 10 points if end of match coincides

- e.g.

```
(alingval '|S@cuisine|
'((NAME M@LIMOUSINE 26)(NAME M@VINE 19)
  (NAME M@MORPHINE 19)
  (NAME M@ENGINE-GOVERNOR 19)
  … 120 more …
))
```



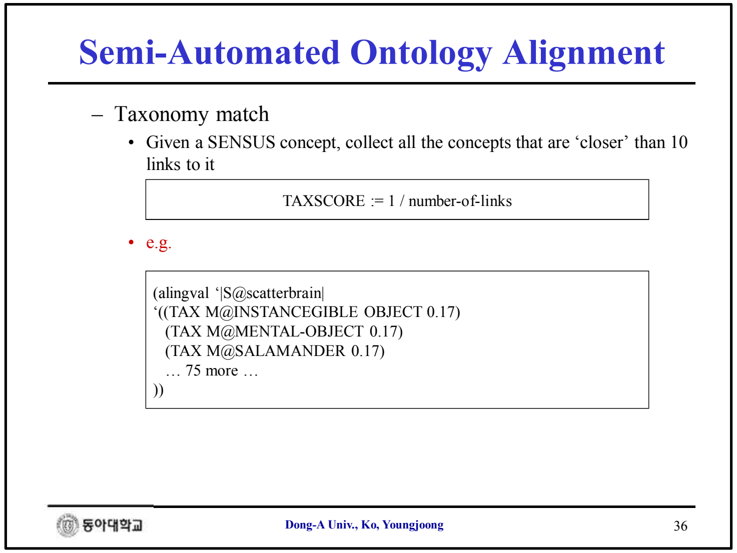
Semi-Automated Ontology Alignment

- Definition match
  - Both definitions are separated into individual word
  - Demorphed
  - Function words and other stop words are removed

DEFSCORE := (Shared(D1,D2)/min(D1,D2))*Shared(D1,D2)

- e.g.

```
(alingval '|S@cuisine|
'((DEF M@KITCHEN (0.62 5 3.12))
  (DEF M@CHEESE (0.62 5 3.12))
  (DEF M@FOODSTUFF (0.62 5 3.12))
  … 9 more …
))
```



Semi-Automated Ontology Alignment

- Taxonomy match
  - Given a SENSUS concept, collect all the concepts that are 'closer' than 10 links to it

TAXSCORE := 1 / number-of-links

- e.g.

```
(alingval '|S@scatterbrain|
'((TAX M@INSTANCEGIBLE OBJECT 0.17)
  (TAX M@MENTAL-OBJECT 0.17)
  (TAX M@SALAMANDER 0.17)
  … 75 more …
))
```

Semi-Automated Ontology Alignment

- Combination Function
  - Must increase with increasing values of NAME, DEF, and TAX
  - Must normalize the heuristics scores
  - Must mitigate the NAME scores' tendency to grow large quickly
  - Must mitigate the TAX scores' tendency to diminish quickly

```
SCORE := sqrt(NAMESCORE) * DEFSCORE * (10 * TAXSCORE)
```

- If NAMESCORE or DEFSCORE are zero, they are replaced by 1
- If TAXSCORE is 0, it is replaced by 0.01

Semi-Automated Ontology Alignment

- Experimental Results
  - Settings
    - MIKRO: 4,790 concepts
    - SENSUS: 6,768 concepts

cutoff	1.4	10	7.8	12	15
New heur	NAME, DEF,TAX	TAX	TAX	TAX	TAX
Total	187	151	170	218	241
Correct	73	11	18	36	106
Near	51	92	51	60	2
wrong	63	48	101	122	39

Semi-Automated Ontology Alignment

- Final Results

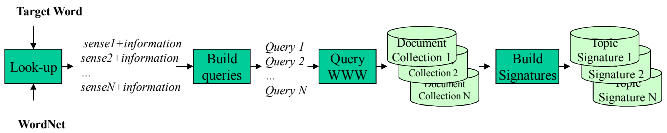
Suggestions	883 (13%) of the portion of SENSUS under consideration
Correct	244 (27.6%)
Nearly Correct	256 (30.0%)
Incorrect	383 (43.4%)

Enriching Ontologies Using WWW

- Building repositories (Ontology)
  - Need huge efforts and investments
  - Unclear results (e.g. WordNet)
    - the lack of relations between topically related concepts
      - e.g. no link between pairs like *bat-baseball, fork-dinner, chicken-farm, etc.*
    - The proliferation of word senses
      - e.g. line has 32 word senses

Building Topic Signatures

- Topic Signatures
  - A list of closely related words for each concept in WordNet
    - e.g. word senses for the noun ‘waiter’
      1. waiter, server – a person whose occupation is to serve at table
         - list: restaurant, menu, waitress, dinner, lunch, counter, etc.
      2. waiter – a person who waits or awaits
         - list: hospital, station, airport, boyfriend, girlfriend, cigarette, etc.



Building the Queries

- The Original Goal of this Step
  - Retrieve from the web all documents related to an ontology concept

- The Queries from the Information in the Ontology

(x AND (cueword$_{1,i}$ OR cueword$_{2,i}$ ...) *#target concept*
AND NOT (cueword$_{1,j}$ OR cueword$_{2,j}$ ... OR *#remaining concepts*
cueword$_{1,k}$ OR cueword$_{2,k}$ ...)

  - Deciding which of the cue-words to use
    - Nouns in the definition are preferable
    - Monosemous cue-words are more valuable
    - Synonyms

Building the Queries

- Information for Sense 1 of ‘*boy*’

synonyms	*male child, child*
gloss	*a youthful male person*
hypernyms	*male, male person*
hyponyms	*alter boy, ball boy, bat boy, cub, lad, laddie, sonny, sonny boy, boy scout, farm boy, plowboy,...*
coordinate sisters	*chap, fello, lad, gent, fella, blighter, cuss, foster, brother, male child, boy, child, man, adult male, ...*

(boy AND (‘alter boy’ OR ‘ball boy’ OR ‘male person’ ...)
AND NOT (‘man’... OR ‘broth of a boy’ OR *# sense 2*
‘son’ OR... OR ‘mama’s boy’ OR *# sense 3*
‘nigger’ OR ... OR ‘black’) *# sense 4*

Build Topic Signatures

- Search the Internet
  - Using AltaVista search engine

- Topic Signature
  - Construct the vector of words from text collection of each word sense
    - Formed with all words and their frequencies
  - Signature function

$$w_{i,j} = \frac{(freq_{i,j} - m_{i,j})}{m_{i,j}},\ m_{i,j} = \frac{\sum_i freq_{i,j} \sum_j freq_{i.j}}{\sum_{i,j} freq_{i,j}}$$

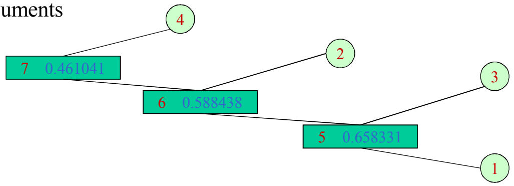
Apply Signatures For WSD

- The Goal of this Experiment
  - To evaluate the automatically constructed topic signatures
  - Apply very simple and straightforward algorithm
    - Collect 50 words from target word's context
    - Sum these important weights in the corresponding topic signature

Word	#senses	#occ	Random	Signatures
Accident	2	12	0.50	0.50
Action	8	130	0.12	0.02
Age	3	104	0.33	0.60
Amount	4	103	0.25	0.50
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
World	8	210	0.12	0.34
Overall	83	2444	***0.28***	***0.41***

Clustering Word Senses

- The Problem
  - WordNet has very fine distinctions between word senses
    - e.g. 'boy'
      - *1: male child, boy, child – a youthful male person*
      - *2: boy – a friendly informal reference to a grown man*
      - *3: son, boy – a male human offspring*
      - *4: boy – offensive term for Black man*
  - Binary Hierarchical Clustering directly on the retrieved documents



Guarino, N., 1998, "Formal Ontology and Information Systems," Proceedings of FOIS'98, Trento, Italy, 6-8.

Remi Zajac., "Engineering Ontologies vs. Linguistic Ontologies for Web Applications,"

Mahesh., K., 1996, "Ontology Development for Machine Translation: Ideology and Methodology," Memoranda in Computer and Cognitive Science MCCS-96-292, Computer Research Laboratory, New Mexico State University.

Knight, K. and Luk, S., 1994, "Building a Large-Scale Knowledge Base for Machine Translation," Proceedings of the American Association of Artificial Intelligence AAAI-94, Seattle, WA

Hovy, E., 2005, "Methodologies for the Reliable Construction of Ontological Knowledge," Proceedings of the 13th Annual International Conference on Conceptual Structures (ICCS 2005) pp. 91-106, LNCS, 3596, 2005.

Agirre, E., Ansa, O., Hovy, E., and Martinez, D. 2000, "Enriching Very Large Ontologies Using the WWW," Proceedings of ECAI Workshop on Ontology Learning, Berlin, August 2000.

Philpot, A.G., Fleischman, M., and Hovy, E., "Semi-automatic Construction of a General Purpose Ontology," Proceedings of the International Lisp Conference, New York, October, 2003.

Hovy, E. "Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses," *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*. Granada, Sp, 1998.

Agirre, E., Arregi, X, Artola, X., Diax de Ilarazza, A., Sarasola, K., "Conceptual Distance and Automatic Spelling Correction," Proceedings of the Workshop on Computational Linguistics for Speech and Handwriting Recognition, Leeds, England.