NLPwin – an introduction

Lucy Vanderwende[1]

Microsoft Research
One Microsoft Way,
Redmond WA 98052
U.S.A.

NLPwin is a software project at Microsoft Research that aims to provide Natural Language Processing tools for Windows (hence, NLPwin). The project was started in 1991, just as Microsoft inaugurated the Microsoft Research group; while active development of NLPwin continued through 2002, it is still being updated regularly, primarily in service of Machine Translation.[2]

NLPwin was and is still being used in a number of Microsoft products, among which the Index Server (1992-3), Word Grammar Checker (parsing every sentence to logical form since 1996), the English Query feature for SQL Server (SQL Server 1998 - 2000), natural language query interface for Encarta (1999, 2000), Intellishrink (2000), and of course, Bing Translator.

Since we knew that we were developing NLPwin in part to support a grammar checker, the NLPwin grammar is designed to be broad-coverage (i.e., not domain-specific) and robust, in particular, robust to grammar errors. While most grammars are learned from data annotated on the PennTreeBank (Marcus et al., 1993), it is interesting to consider that such grammars may not be able to parse ungrammatical or fragmented grammar, since those grammars have no training data for such input. The NLPwin grammar produces a parse for any input and if no spanning parse can be assigned, it creates a "fitted" parse, combining the largest constituents that it was able to construct.

The NLP rainbow: we envisioned that with ever more sophisticated analysis capabilities, it would be possible to create applications of a wide variety. As you can see below, the generation component was not well developed and we postulated NL applications for generation much as one hopes for a pot of gold at the end of the rainbow. Our first MT models transferred at the semantic level (papers through 2002), while today, our MT transfers primarily at the syntactic level, using a mixture of syntax-based and phrase-based models.

[1] On behalf of everyone who contributed to the development on NLPwin

[2] A near-complete list of publications that describe the creation and definition of the NLPwin system as well as describe multiple uses of the NLPwin system is included in this techreport.

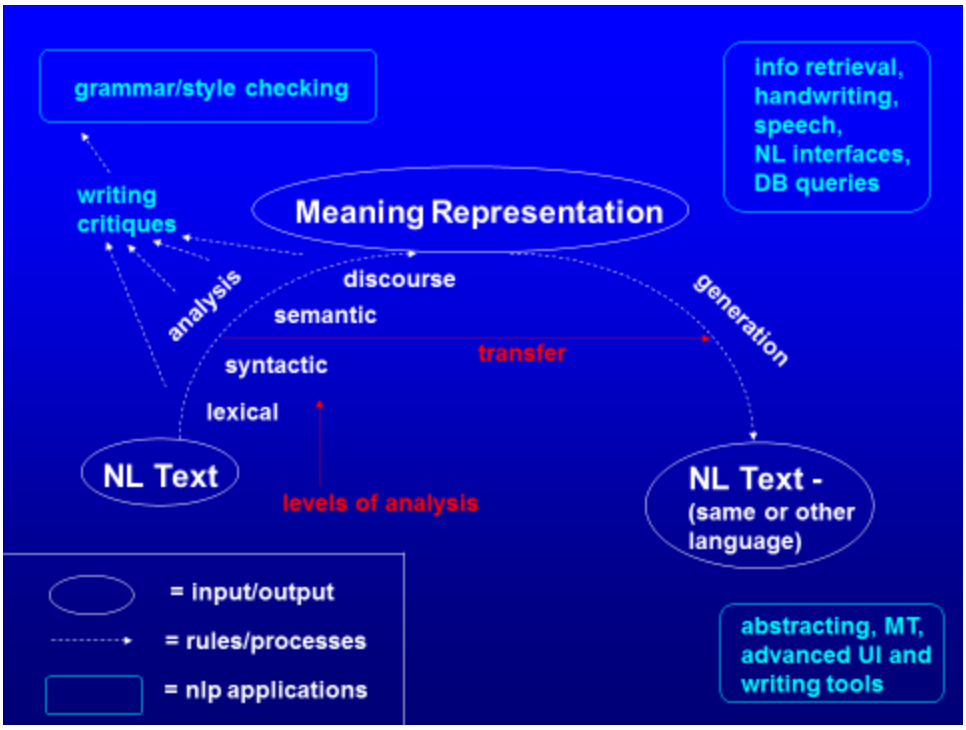


Figure 1: The NLP rainbow (1991), our original vision for NLP components needed and applications possible

The architecture follows a pipeline approach, as shown in Figure 2, where each component provides additional layers of analysis/annotation of the input data. We designed the system to be relatively knowledge-poor in the beginning, while making use of richer and richer data sources as the need for more semantic information increased; one of our goals of this architecture is to preserve ambiguity until we either needed to resolve that ambiguity or the data resources existed to allow the resolution. Thus, the syntactic analysis proceeds in two steps: the syntactic sketch (which today might be described as a packed forest) and the syntactic portrait, where we "unpack" the forest and construct a constituent level of analysis which is syntactic, but also semantically valid. The constituency tree continues to be refined even during Logical Form processing as more global information can be brought to bear.

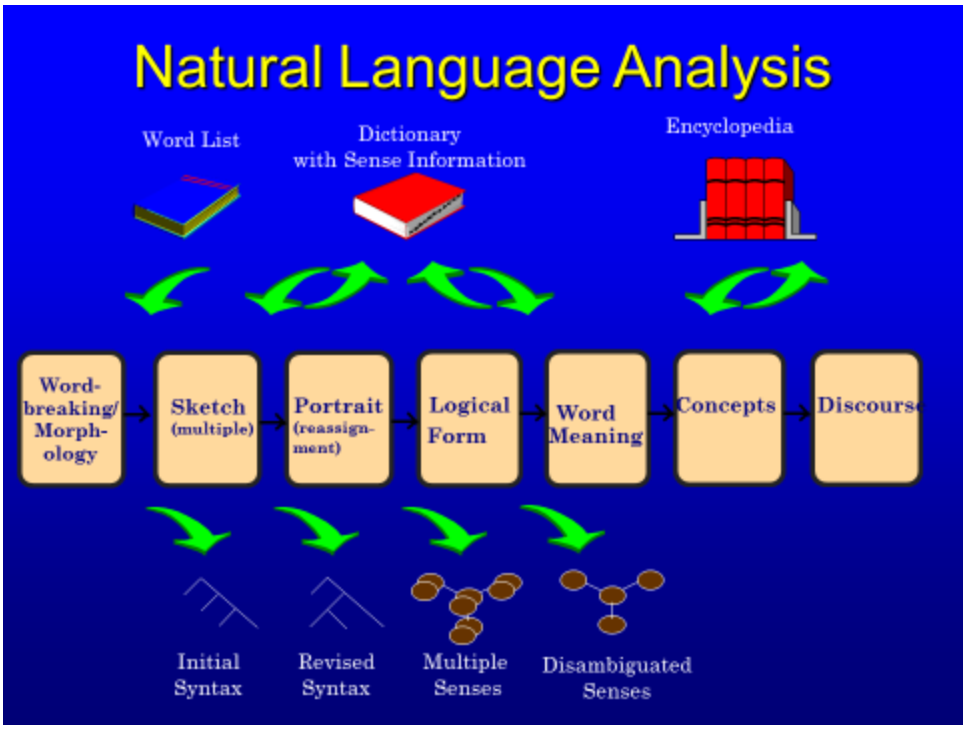


Figure 2: The NLPwin components and a schematic of their output representation.

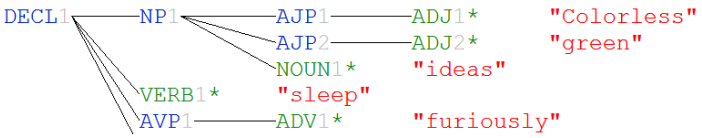
A few points are worth making about the parser (a term which loosely combines the morphology, sketch and portrait modules). First, the parser is comprised of human authored rules. This will cause incredulity among those who are only familiar with machine-learned parsers that have been trained on the PennTreeBank. It should be kept in mind that the NLPwin parser was constructed before the first parser was trained on the PennTreeBank, that the parser had to be fast (to support the grammar checker) and that grammar rule-writing was the norm pre-PennTreeBank grammars. Furthermore, the grammarian tasked with writing rules was supported by a sophisticated array of NLP developer tools (created by George Heidorn) (see Suzuki, 2002), much as a programmer is now supported in Visual Studio, where grammar rules can be run to and from specific points in the code, variables can be changed interactively for exploration purposes, and most importantly, the developer environment supported running a suite of test files with interfaces for the grammarian to update the target files with improved parses. Secondly, the lead grammarian, Karen Jensen, broke with the implicit tradition where the constituent structure is implied by application of the parsing rules[3]. Jensen observed that binary rules are required to handle even common language phenomena such as free word order, and adverbial and prepositional phrase placement. Thus, in NLPwin, we use binary rules in an augmented phrase structure grammar formalism (APSG), computing the phrase structure as part of the actions of the rules, thereby creating nodes with unbounded modifiers, while maintaining binary rules, illustrated in Figure 3.

[3] see Karen Jensen. 1987. Binary rules and non-binary trees: Breaking down the concept of phrase structure. In Mathematics of language, ed. A. Manaster-Ramer, 65-86. Amsterdam: John Benjamins Pub.Co.

Rules not isomorphic to tree structure

Computed Tree

Note: simplified structure with only single NP node pre-modifying the head VERB1, regardless how many adjectives are present.



Derivation Tree

Note: structure strictly follows rule application, with two NP nodes (NP2 and NP3), one for each "NP with Adj to the left" rule application.

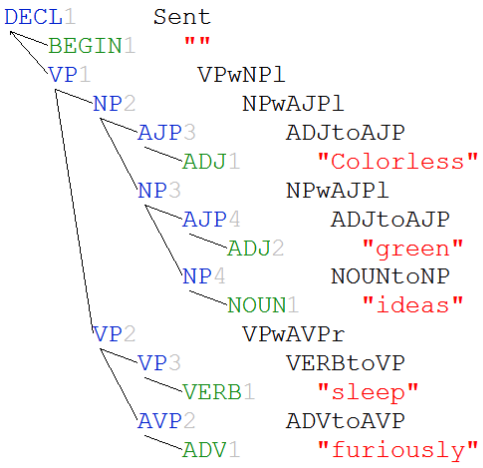


Figure 3: The derivation tree displays the history of rule application, while the computed tree provides a useful visualization of phrase structure

Another important aspect of NLPwin is that it is the record structure, not the trees, that is the fundamental output of the analysis component (shown in Figure 4). Trees are merely a convenient form of display, using only 5 of the many attributes that make up the representation of the analysis (premodifiers (PRMODS), HEAD, postmodifiers (PSMODS), segment-type (SEGTYPE), and string value. Here is the record, a collection of attributes and values, for the node DECL1:

Records are the primary goal of analysis

Blue are the attributes displayed in the tree,
Green are some of the many useful attributes computed and consulted during analysis

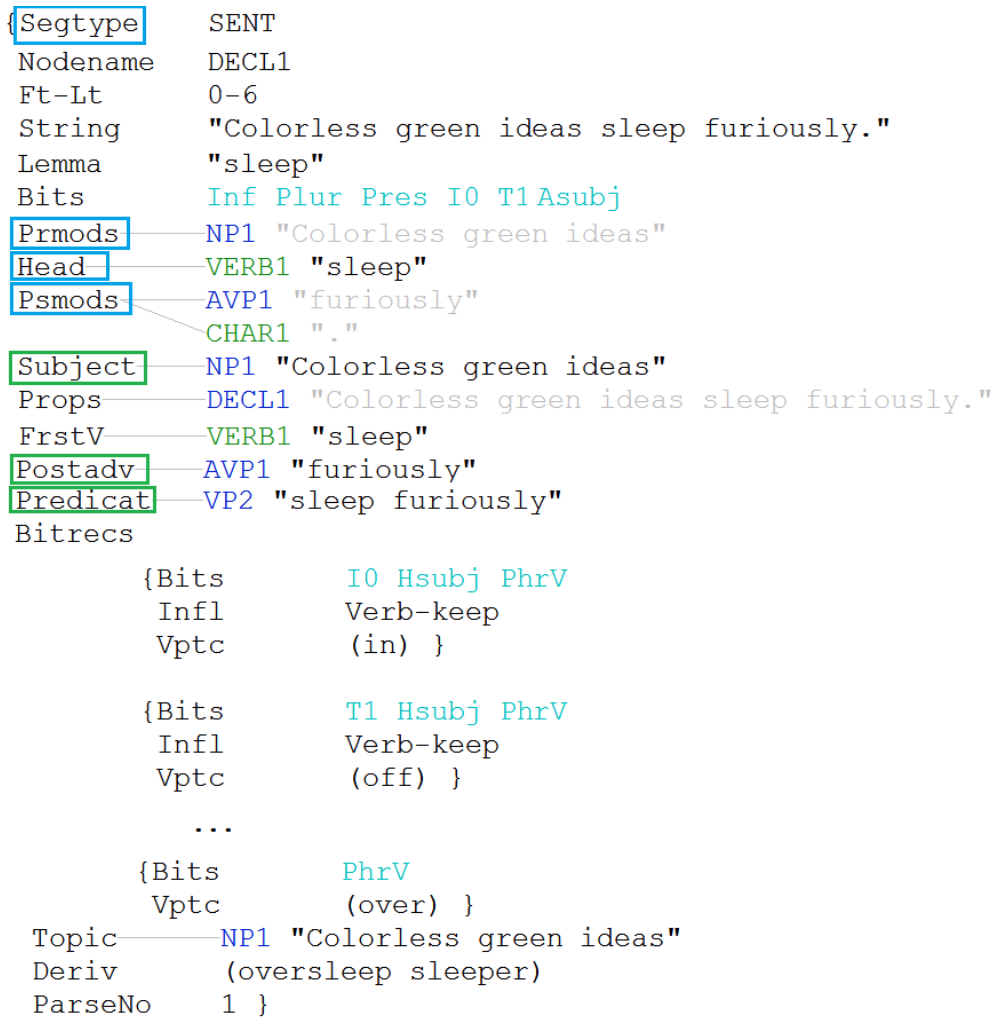


Figure 4. The record structure of any constituent is the heart of the NLPwin analysis

Once the basic shape of the constituency tree has been determined, it is possible to compute what the Logical Form is. The goal of Logical Form is twofold: to compute the predicate-argument structure for each clause ("who did what to whom when where and how?") and to normalize differing syntactic realizations of what can be considered the same "meaning". In so doing, concepts that are possibly distant in the sentence and in the constituent structure can be brought together, in large part because the Logical Form is represented as a graph, where linear order is no longer primary. The Logical Form is a directed, labeled graph, where arcs are labeled with those relations that are defined to be semantic and surface words that convey syntactic information only are represented not as nodes in the graph but

rather as annotations on the nodes, preserving their syntactic information (not shown in the graph representation below). Consider the following Logical Form:

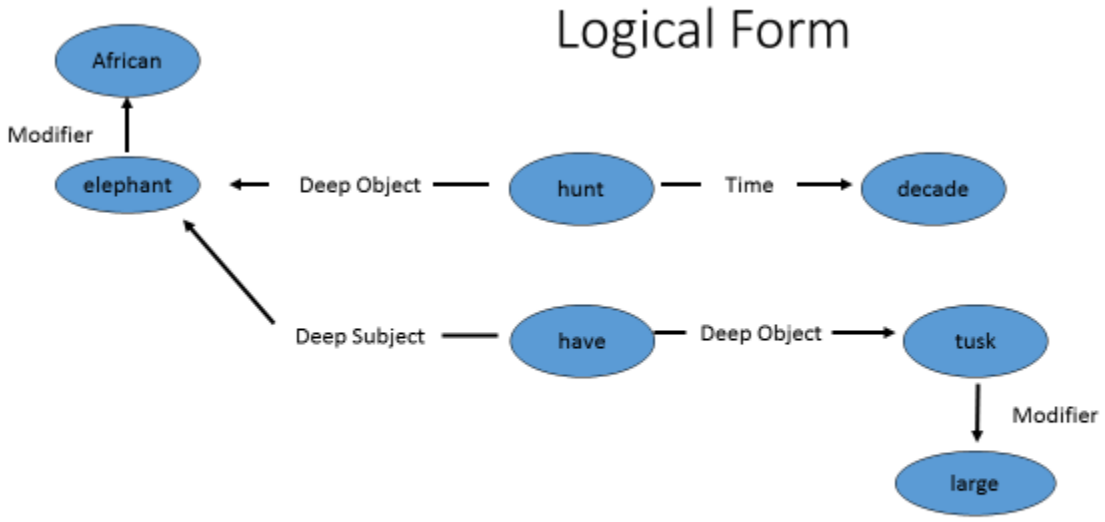


Figure 5. A Logical Form example

The Logical Form graph in Figure 5 represents the direct connection between "elephants" and "have", which is interrupted by a relative clause at the surface syntax. Moreover, in analyzing the relative clause, the Logical Form has performed two operations: Logical Form normalizes the passive construction as well as assigns the referent of the relative pronoun "which". Other operations commonly performed by Logical Form include (but are not limited to): unbounded dependencies, functional control, indirect object paraphrase, assigning modifiers.

Figure 5 also demonstrates some of the shortcomings of Logical Form: 1) should "have" be a concept node in this graph or should it be interpreted as an arc labeled Part between "elephant" and "tusk"? More generally: what should the inventory of relation labels be, and how should that inventory be determined? And 2) should we infer from this sentence only that "African elephants have been hunted" and that "African elephants have large tusks", or can we infer that "elephants have been hunted" and that they happen to be "African elephants". Deciding this question of scoping was postponed till discourse processing[4], when such questions may be addressed, and Logical Form does not represent the ambiguity in scoping.

During development of the NLPwin pipeline (see Figure 2), we considered that there would be a separate component determining word senses following the syntactic analysis of the input. This component was meant to select and/or collate lexical information from multiple dictionaries to represent and expand the lexical meaning of each content word. This view on Word Sense

[4] In fact, the NLPwin system has not (yet) addressed this issue till today.

Disambiguation (WSD) was in contrast to the then-nascent interest in WSD in the academic community, which formulated the WSD task as selecting one sense of a fixed inventory of word senses as being correct. Our primary objection to this formulation is that any fixed inventory will necessarily not be sufficient as the foundation for a broad-coverage grammar (see Dolan, Vanderwende and Richardson, 2000 but also Palmer et al. 2004, Snow et al., 2007, i.a.). For similar reasons, we elected to abandon the pursuit of assigning Word Senses in NLPwin as well. Today, the field has made great strides in exploring a more flexible notion of lexical meaning with the advent of vector space, which would be promising to combine with the output of this parser.

While we did not view Word Sense Disambiguation as a separate task, we did design our parser and subsequent components to make use of ever richer lexical information. The sketch grammar relies on the subcategorization frames and other syntactic-semantic codes available from two dictionaries: Longman Dictionary of Contemporary English (LDOCE) and American Heritage Dictionary, 3rd edition, for which Microsoft had acquired the digital rights. LDOCE in particular provides rich lexical information[5] that facilitates the construction of Logical Form. Such codes, rich as they are, do not support full semantic processing as is necessary when, for example, determining the correct attachment of prepositional phrases or nominal co-reference. The question was: is it possible to acquire such semantic knowledge automatically, in order to support a broad-coverage parser?

In the early to mid-90s, there was considerable interest in mining dictionaries and other reference works for semantic information broadly-speaking. For this reason, we envisioned that where lexical information was not sufficient to support the decisions that needed to be made in the Portrait component, we would acquire such information in machine readable reference works.

At the time, few broad-coverage parsers were available so the main thrust was to develop string patterns (regexes) that could be used to identify specific types of semantic information; Hearst (1992) describes the use of such patterns for the acquisition of Hypernymy (is-a terms). Alshawi (1989) parses dictionary definitions using a grammar especially designed for that dictionary ("Longmanese"). We encountered two concerns about using this approach: first, as the need for greater recall increases, writing and refining string patterns becomes more and more complex, in the limit, approaching the complexity of full-grammar writing and so straying far from the straightforward string patterns you started with, and second, when extracting semantic relations beyond Hypernymy, we found string patterns to be insufficient (see Montemagni and Vanderwende 1992).

Instead, we proposed to parse the dictionary text using the linguistic components already developed, Sketch, Portrait and Logical Form, ensuring access to robust parsing, in order to bootstrap the knowledge acquisition of the semantic information needed to improve the Portrait. This bootstrapping is possible because some linguistic expressions are unambiguous, and so, at each iteration, we can extract from unambiguous text to improve the parsing of ambiguous text (see Vanderwende 1995).

[5] The LDOCE box codes, for instance, provide information on type restrictions and the arguments for verbs. In LDOCE, "persuade" is marked ObjC, indicating, that "persuade" has Object Control (i.e. that the object of "persuade" is understood to be the subject of the verb complement). Thus, it is possible to construct a Logical Form with "John" as the subject of "go to the library" from the input sentence: "I persuaded John to go to the library", while for the input sentence "I promised John to go to the library", the Logical Form is constructed with "I" as the subject of "go to the library".

As each definition in the dictionary and on-line encyclopedia was being processed and the semantic information was being stored for access by Portrait, a picture emerged from connecting all of the graph fragments. When viewed as a database rather than a look-up table (which is how people use dictionaries), the graph fragments are connected and interesting paths/inferences emerge. To enrich the data further, we then took the step of viewing each graph fragment from the perspective of each content node. Imagine looking at the graph as a mobile and picking it up at each of the objects in turn - the nodes under the object remain the same, but the nodes above that object become inverted (illustrated in Figure 6). For example, for the definition of **elephant**: :an animal with ivory tusks", MindNet stores not only the graph fragment "elephant PART (tusk MATR ivory)" but also "tusk PART-OF elephant" and "ivory MATR-OF tusk"[6].

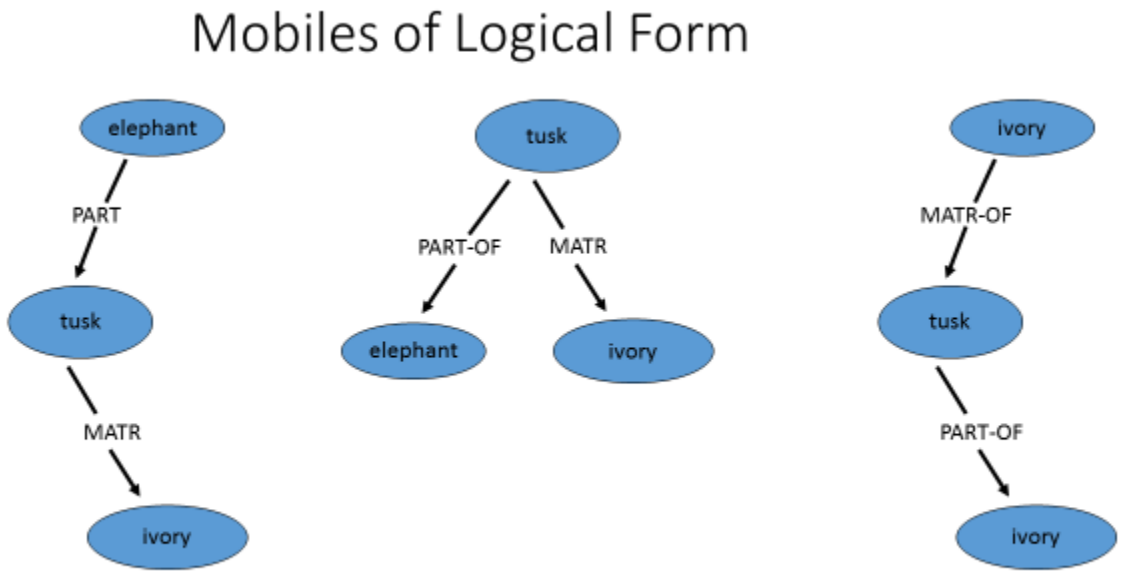


Figure 6. Logical Form and its inversions

We called this collection of intersecting graphs MindNet. Figure 7 reflects the picture we saw for the word "bird" when looking at all of the pieces of information that were automatically gleaned from dictionary text:

[6] The algorithm of course also identifies the relation "elephant HYPERNYM animal", but, in dictionary processing, the information extracted from the differentiae of the definition (the specifications on the hypernym), are true of the word being defined rather than true of the hypernym, and so we do not extract that "animals have tusks" but rather that "elephants have tusks".

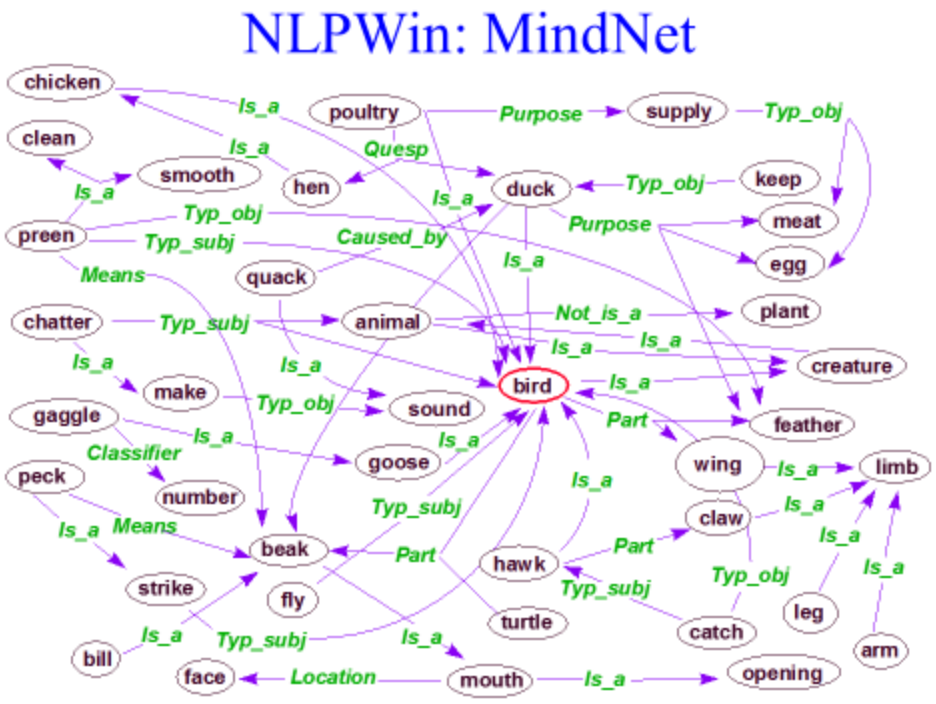


Figure 7: A fragment NLPwin MindNet, centered on the word "bird"

As a person using only the dictionary, it would be very difficult to construct a list of all the different types of birds, all of the parts of a bird, all of the places that a bird may be found, or the types of actions that a bird may do. But by converting the dictionary to a database, and inverting all the semantic relations as shown in Figure 6, MindNet contains rich semantic information for any concept that occurs in text, esp. because it is produced by automated methods using a broad-coverage grammar, a grammar that parses fragments as well as it parses complete grammatical input.

We computed a similarity metric for MindNet (Richardson, Dolan and Vanderwende, 1998) by using Roget's Thesaurus as annotated training data. Given a pair of words from Roget, we computed all paths in MindNet between those synonyms and then observed how often path patterns occur (patterns of relation types with and without the specific concepts linking those relations types). Thus, we learn that if X and Y are connected using the path pattern: (**X** - Hypernym - *z* - HypernymOf **Y)** or (**X** - HasObject - *z* - ObjectOf - **Y**), that X and Y are deemed to be similar with high weight. We can then query arbitrary word pairs for their similarity, finding that "gold" and "zinc" are similar, while "gold" and "bicycle" are not.

A priori, there is no reason that MindNet cannot be produced from text other than dictionary or encyclopedia text. Indeed, if MindNet was being developed today, we would aim to automatically acquire semantic information from the web. The notable engineering concern is processing time, though the availability of massively parallel web services mitigates that concern in large part. The other notable concern is to establish the veracity of the source material (part of IBM Watson's success in the Jeopardy game can be attributed to careful selection of its information sources (Chu-Carroll et al., 2012). Even in the case where sources are equally trustworthy, what should happen to (apparent) contradictions? Weights computed for specific pieces of the knowledge graph can be used to balance how frequently that information is encountered, but the source itself should also be considered in the weight scheme.

Moreover, MindNet is not simply a database of triples; we preserve the context from which the semantic relations were extracted, and so in theory, we could resolve apparent contradictions by taking context into account. We did not encounter these concerns as MindNet has only been computed from sources that are categorically true (dictionaries and encyclopedias), but these concerns should be addressed going forward with knowledge acquisition from the web.

The original intent, as shown in Figure 2, was to reduce paraphrases to a canonical representation in a module that we tentatively named “Concepts”, though "Concept Detection" would have been more descriptive. As with Word Sense Disambiguation, we abandoned this module as we were dissatisfied with the underlying assumption that one representation of a concept or complex event would be primary over others, while in reality, both expressions are equivalent; equivalence should be fluid and allow to vary depending on the need of the application. Here again, we believe that the current research which aims to represent parse fragments in vector space is a promising approach, while emphasizing that it is essential to take the parse and logical form structure into account.

Finally, a few words about the generation grammar (shown on the right hand side of the rainbow in Figure 1). In NLPwin, we developed two types of generation grammars: rule-based generation components (including those that shipped with Microsoft Word to enable the re-write of passive to active, e.g.) and Amalgam, a set of machine-learned generation modules. Both types of generation grammars were used in production for Machine Translation.

In summary ...

We’ve described some of the aspects of the NLPwin project at Microsoft Research[7]. The lexical and syntactic processing components are designed to be broad-coverage and robust to grammatical errors, allowing for parses to be constructed for fragmented, ungrammatical as well as grammatical inputs. These components are largely rule-based grammars, making use of rich lexical and semantic resources derived from online dictionaries. The output of the parsing component, a tree analysis, is converted to a graph-based representation called Logical Form. The goal of Logical Form is to compute the predicate-argument structure for each clause and to normalize differing syntactic realizations of what can be considered the same “meaning”. In so doing, the distance between concepts reflects the semantic distance and no longer the linear distance in the surface realization, bringing related concepts closer together than they might appear at the surface. MindNet is the automatic construction of the database of connected Logical Forms. When reference resources are the source text for MindNet, MindNet can be viewed as a traditional Knowledge Acquisition method and object, but when MindNet is constructed by processing arbitrary text input, MindNet represents a global representation of all the Logical Forms of that text which allows the browsing of the concepts and their semantic connections in that text. In fact, MindNet was considered most compelling as a means for browsing and exploring specific relations mined from a text collection.

[7] At the time of this writing (2014) NLPwin is considered a mature system, with only limited development of the eneration and logical form components.

NLPwin Techreport References

Hiyan Alshawi. 1987. Processing Dictionary Definitions with Phrasal Pattern Hierarchies. In Computational Linguistics, Volume 13, Numbers 3-4, July-December 1987

Jennifer Chu-Carroll, J Fan, N Schlaefer, and Wlodek Zadrozny. 2012. Textual resource acquisition and engineering. In IBM Journal of Research and Development 56(3-4), 2012.

William Dolan, Lucy Vanderwende, and Stephen D. Richardson. 2000. Polysemy in a Broad-Coverage Natural Language Processing System. In Polysemy: Theoretical and Computational Approaches. Eds. Yael Ravin and Claudia Leacock. Oxford University Press, July 2000.

Marti Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992.

Karen Jensen. 1987. Binary rules and non-binary trees: Breaking down the concept of phrase structure. In Mathematics of language, ed. A. Manaster-Ramer, 65-86. Amsterdam: John Benjamins Pub.Co.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. In Computational Linguistics Journal, 19:313-330.

Simonetta Montemagni and Lucy Vanderwende. 1992. Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries. In Proceedings of the Fourteenth International Conference on Computational Linguistics, COLING-1992

Martha Palmer, Olga Babko-Malaya, Hoa Trang Dang. 2004. Different Sense Granularities for Different Applications. In Proceedings of Workshop on Scalable Natural Language Understanding.

Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In Proceedings of COLING-ACL 1998

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In Proceedings of EMNLP.

Hisami Suzuki. 2002. A development environment for large-scale multi-lingual parsing systems. In Proceedings of the 2002 workshop on Grammar engineering and evaluation - Volume 15, Pages 1-7.

Lucy Vanderwende. 1995. Ambiguity in the Acquisition of Lexical Information. In AAAI Symposium on Representation and Acquisition of Lexical Knowledge*:* TR SS-95-01, AAAI, 1995

NLPwin Team Members

English & core development:

Karen Jensen, George Heidorn, Stephen D. Richardson, Diana Peterson, Lucy Vanderwende, Joseph Pentheroudakis, Bill Dolan, Deborah Coughlin, Lee Schwartz, Simon Corston Oliver, Eric Ringger, Rich Campbell, Arul Menezes, Chris Quirk;

French: Martine Pettenaro, Jessie Pinkham, Martine Smets;

Spanish: Marisa Jimenez, Carmen Lozano, Maite Melero;

German: Michael Gamon, Tom Reutter;

Japanese: Takako Aikawa, Chris Brockett, Hisami Suzuki;

Chinese: Terrence Peng, Andi Wu, Jiang Zixin;

Korean: Jee Eun Kim, Kong Joo Lee

NLPwin publications[8]

Papers that describe NLPwin

Development Environment

Hisami Suzuki. 2002. A development environment for large-scale multi-lingual parsing systems. In Proceedings of the 2002 workshop on Grammar engineering and evaluation - Volume 15, Pages 1-7.

Morphology

Joseph Pentheroudakis and Lucy Vanderwende, Automatically Identifying Morphological Relations in Machine-Readable Dictionaries. 1993. In Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research.

Syntax

Jensen, Karen, George E. Heidorn and Stephen D. Richardson (eds.). 1993. **Natural Language Processing: The PLNLP approach**. Kluwer: Boston.

NOTE: While this is not a reference for the work done at Microsoft, the PLNLP approach provides a good overview of the motivation and design of the syntax system, as well as a number of other key components of the complete NLP system.

Stephen D. Richardson. 1994. Bootstrapping Statistical Processing into a Rule-based Natural Language Parser. In Proceedings of the The Balancing Act Workshop: Combining Symbolic and Statistical Approaches to Language, sponsored by ACL.

Michael Gamon and Tom Reutter. 1997. The Analysis of German Separable Prefix Verbs in the Microsoft Natural Language Processing System. Microsoft Research Technical Report, MSR-TR-97-15, September 1997

Michael Gamon, Carmen Lozano, Jessie Pinkham, and Tom Reutter. 1997. Practical Experience with Grammar Sharing in Multilingual NLP, no. MSR-TR-97-16

Michael Gamon, Carmen Lozano, Jessie Pinkham, and Tom Reutter. 1997. From Research to Commercial Applications: Making NLP Work in Practice. In Proceedings of the ACL Workshop "From Research to Commercial Applications: Making NLP Work in Practice"

Takako Aikawa, Chris Quirk, and Lee Schwartz. 2003. Learning prepositional attachment from sentence aligned bilingual corpora, Association for Machine Translation in the Americas.

Lee Schwartz; Takako Aikawa. 2004. Multilingual Corpus-based Approach to the Resolution of English –ing. In Proceedings of LREC.

Logical Form

[8] Please contact lucyv@microsoft.com if you have any additions to this bibliography to suggest or if you wish to suggest any corrections.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In Proceedings of the 15th International Conference on Computational Linguistics, Volume 2.
Lucy Vanderwende. 1996. The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries, PhD thesis, Georgetown University, Microsoft Research Technical Report, no. MSR-TR-95-57, October 1996
Richard Campbell and Hisami Suzuki. 2002. **Language-Neutral Syntax: An Overview.** Microsoft Research Technical Report, MSR-TR-2002-76
Richard Campbell. 2002. Computation of modifier scope in NP by a language-neutral method. In Proceedings of the 19th International Conference on Computational Linguistics, COLING-2002.
Richard Campbell and Hisami Suzuki. 2002. Language-Neutral Representation of Syntactic Structure. In Proceedings of the First International Workshop on Scalable Natural Language Understanding (SCANALU 2002), Heidelberg, Germany
Richard Campbell, Takako Aikawa, Zixin Jiang, Carmen Lozano, Maite Melero and Andi Wu. 2002. A language neutral representation of temporal information. In LREC 2002 Workshop Proceedings: Annotation Standards for Temporal Information in Natural Language. 13-21.
Richard Campbell and Eric Ringger. 2004. Converting Treebank Annotations to Language Neutral Syntax, In Proceedings of LREC.
Richard Campbell. 2004. Using linguistic principles to recover empty categories. In Proceedings of ACL.

Word Sense Disambiguation

William B. Dolan. 1994. Word sense ambiguation: clustering related senses. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, 5-9 August 1994, Kyoto, Japan, 712-716.
William Dolan, Lucy Vanderwende, and Stephen D. Richardson. 2000. Polysemy in a Broad-Coverage Natural Language Processing System. In **Polysemy: Theoretical and Computational Approaches**. Eds. Yael Ravin and Claudia Leacock. Oxford University Press, July 2000.

Discourse

Simon Corston-Oliver. 1998. Beyond string matching and cue phrases. In Proceedings of AAAI 98 Spring Symposium on Intelligent Text Summarization.
Simon H. Corston-Oliver. 1998. Identifying the Linguistic Correlates of Rhetorical Relations. In Discourse Relations and Discourse Markers workshop at COLING-ACL98.
Simon H. Corston-Oliver. 2000. Using decision trees to select the gran natical relation of a noun phrase. In Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue, ACL.

MindNet – Automatic construction of a Knowledge Base

Simonetta Montemagni and Lucy Vanderwende. 1992. Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries, in Proceedings of the Fourteenth International Conference on Computational Linguistics, COLING-1992
William Dolan, Stephen D. Richardson, and Lucy Vanderwende. 1993. Automatically Deriving Structured Knowledge Bases From On-Line Dictionaries, no. MSR-TR-93-07, May 1993
William Dolan, Stephen D. Richardson, and Lucy Vanderwende. 1993. Combining Dictionary-Based and Example-Based Methods for Natural Language Analysis , no. MSR-TR-93-08, June 1993
Lucy Vanderwende. 1995. Ambiguity in the Acquisition of Lexical Information. In AAAI Symposium on Representation and Acquisition of Lexical Knowledge: TR SS-95-01, AAAI, 1995

Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In Proceedings of COLING-ACL 1998

Lucy Vanderwende, Gary Kacmarcik, Hisami Suzuki, and Arul Menezes. 2005. MindNet: an automatically-created lexical resource. In Proceedings of the HLT/EMNLP Interactive Demonstrations, October 2005

Generation

Simon Corston-Oliver, Michael Gamon, Eric Ringger, and Robert Moore. 2002. An overview of Amalgam: A machine-learned generation module. In Proceedings of ACL

Michael Gamon, Eric Ringger, and Simon Corston-Oliver. 2002. Amalgam: A machine-learned generation module. Microsoft Research Technical Report no. MSR-TR-2002-57, June 2002

Zhu Zhang, Michael Gamon, Simon Corston-Oliver, Eric Ringger. 2002. Intra-sentence Punctuation Insertion in Natural Language Generation. Microsoft Research Technical Report no. MSR-TR-2002-58

Other languages:

German

Michael Gamon, Eric Ringger, Simon Corston-Oliver, and Robert C. Moore. 2002. Machine-learned contexts for linguistic operations in German sentence realization, In Proceedings of ACL.

Michael Gamon, Eric Ringger, Zhu Zhang, Robert Moore, and Simon Corston-Oliver. 2002. Extraposition: A case study in German sentence realization,. In Proceedings of ACL

French

Martine Smets, Michael Gamon, Simon Corston-Oliver, and Eric Ringger. 2003. The adaptation of a machine-learned sentence realization system to French, In Proceedings of the European chapter of ACL

Martine Smets, Michael Gamon, Simon Corston-Oliver, and Eric Ringger. 2003. French Amalgam: A machine-learned sentence realization system, Association pour le Traitement Automatique des Langues, TALN 2003

Spanish

Melero, M., T. Aikawa and L. Schwartz. 2002. Combining machine learning and rule-based approaches in Spanish and Japanese sentence realization. In Proceedings of the Second International Natural Language Generation Conference

Chinese

Wu, Andi and Zixin Jiang. 1998. Word Segmentation in Sentence Analysis. Microsoft Technical Report MSR-TR-99-10.

Andi Wu, Joseph Pentheroudakis, and Zixin Jiang. 2002. Dynamic Lexical Acquisition in Chinese Sentence Analysis In Project Notes at the International Conference on Computational Linguistics, COLING-2002.

Japanese

Gary Kacmarcik, Chris Brockett and Hisami Suzuki. 2000. Robust Segmentation of Japanese Text into a Lattice for Parsing. In Proceedings of COLING 2000.

Hisami Suzuki, Chris Brockett and Gary Kacmarcik. 2000. Using a Broad-Coverage Parser for Word-Breaking in Japanese. In Proceedings of COLING 2000.

Papers that make use of NLPwin output

Grammar Checker

George E. Heidorn. 2000. Intelligent writing assistance. In **A Handbook of Natural Language Processing**: Techniques and Applications for the Processing of Language as Text. Marcel Dekker, New York. pp. 181-207.

Machine Translation

Michael Gamon, Hisami Suzuki, and Simon Corston-Oliver. 2001. Using Machine Learning for System-Internal Evaluation of Transferred Linguistic Representations, European Association for Machine Translation, January 2001

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation, Association for Computational Linguistics

Arul Menezes and Stephen D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora, Association for Computational Linguistics

William Dolan, Stephen D. Richardson, Arul Menezes, and Monica Corston-Oliver. 2001. Overcoming the customization bottleneck using example-based MT, Association for Computational Linguistics

Stephen D. Richardson, William B. Dolan, Arul Menezes, and Jessie Pinkham. 2001. Achieving commercial quality translation with example-based methods. In Proceedings of MT Summit VIII, Santiago De Compostela, Spain. 293-298.

Richard Campbell, Carmen Lozano, Jessie Pinkham, and Martine Smets. 2002. Machine Translation as a Testbed for Multilingual Analysis. In Proceedings of COLING 2002

Jessie Pinkham and Martine Smets. 2002. Modular MT with a learned bilingual dictionary: rapid deployment of a new language pair. In Proceedings of COLING 2002

Jessie Pinkham and Martine Smets. 2002. Machine Translation without a Bilingual Dictionary. In Proceedings of The 9th Conference on Theoretical and Methodological Issues in Machine Translation.

Chris Brockett, Takako Aikawa, Anthony Aue, Arul Menezes, Chris Quirk, and Hisami Suzuki. 2002. English-Japanese Example-Based Machine Translation Using Abstract Semantic Representations, Proceedings of Coling 2002 workshop on Machine Translation in Asia, at COLING-2002

Martine Smets, Joseph Penteroudakis, and Arul Menezes. 2002. Translation of Verbal Idioms.In Proceedings of the International Workshop on Computational Approaches to Collocations, Colloc-02, Vienna, Austria

Arul Menezes. 2002. Better contextual translation using machine learning. In 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, October 8 – 12, 2002 Proceedings, Springer, Verlag

Martine Smets, Michael Gamon, Jessie Pinkham, Tom Reutter, and Martine Pettanaro. 2003. High quality machine translation using a machine-learned sentence realization component. In Proceedings of the Association for Machine Translation in the Americas

Simon Corston-Oliver and Michael Gamon. 2003. Combining decision trees and transformation-based learning to correct transferred linguistic representations. In Proceedings of the Association for Machine Translation in the Americas

Anthony Aue, Arul Menezes, Robert Moore, Chris Quirk, and Eric Ringger. 2004. Statistical Machine Translation Using Labeled Semantic Dependency Graphs. In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2004). Baltimore, Maryland.

Simon Corston-Oliver and Michael Gamon. 2004. Normalizing German and English Inflectional Morphology to Improve Statistical Word Alignment. In Proceedings of the Association for Machine Translation in the Americas

Chris Quirk, Arul Menezes, and Colin Cherry. 2004. Dependency Tree Translation: Syntactically Informed Phrasal SMT, no. MSR-TR-2004-113, November 2004

Eric Ringger, Michael Gamon, Robert C. Moore, David Rojas, Martine Smets and Simon Corston-Oliver. 2004. Linguistically informed statistical models of constituent structure for ordering in sentence realization. In Proceedings of the 20th Internation Conference on Computational Linguistics.

Donghui Feng, Yajuan Lü, Ming Zhou. 2004. A New Approach for English-Chinese Named Entity Alignment. In Proceedings of EMNLP-2004

Yajuan Lü and Ming Zhou. 2004. Collocation Translation Acquisition Using Monolingual Corpora. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

Arul Menezes and Chris Quirk. 2005.Microsoft Research Treelet Translation System: IWSLT Evaluation. In Proceedings of the International Workshop on Spoken Language Translation, October 2005

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In Proceedings of ACL

Arul Menezes and Chris Quirk. 2005. Dependency treelet translation: the convergence of statistical and example-based machine-translation. In Proceedings of the 10th Machine Translation Summit Workshop on Example-Based Machine Translation

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. In Proceedings of the European Association for Machine Translation.

Chris Quirk and Simon Corston-Oliver. 2006. The impact of parse quality on syntactically-informed statistical machine translation. In Proceedings of EMNLP 2006

Xiaodong He, Arul Menezes, Chris Quirk, Anthony Aue, Simon Corston-Oliver, Jianfeng Gao, and Patrick Nguyen. 2006. Microsoft Research Treelet Translation System: NIST MT Evaluation 06, National Institute of Standards and Technology , March 2006

Chris Quirk and Arul Menezes. 2006. Dependency Treelet Translation: The convergence of statistical and example-based machine translation?. In Machine Translation, vol. 20, pp. 43–65, March 2006

Arul Menezes, Kristina Toutanova, and Chris Quirk. 2006. Microsoft research treelet translation system: NAACL 2006 Europarl evaluation. In WMT 2006

Arul Menezes and Chris Quirk. 2007. Using Dependency Order Templates to Improve Generality in Translation. In Proceedings of the Second Workshop on Statistical Machine Translation at ACL 2007

Arul Menezes and Chris Quirk. 2008. Syntactic Models for Structural Word Insertion and Deletion during Translation. In Proceedings of EMNLP 2008

Summarization

Kristina Toutanova, Chris Brockett, Michael Gamon, Jagadeesh Jagarlamundi, Hisami Suzuki, and Lucy Vanderwende. 2007. The Pythy Summarization System: Microsoft Research at DUC 2007. In Proceedings of DUC-20077

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond SumBasic: Task-Focused Summarization with Sentence Simplification and Lexical Expansion. In Information Processing and Management, Volume 43 , Issue 6, pages 1606-1618

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating DUC 2005 using Basic Elements. In Proceedings of the DUC-2005 workshop.

Jurij Leskovec, Natasa Milic-Frayling, Marko Grobelnik. 2005. Impact of Linguistic Analysis on the Semantic Graph Coverage and Learning of Document Extracts. In Proceedings of the National Conference on Artificial Intelligence (AAAI), 2005.

Simon H. Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. 2004. Task-focused summarization of email. In Proceedings of the ACL 2004 Workshop "Text Summarization Branches Out", Barcelona, Spain.

Lucy Vanderwende, Michele Banko, and Arul Menezes. 2004. Event-centric summary generation. In Working notes of the Document Understanding Conference 2004

Jurij Leskovec, Natasa Milic-Frayling, Marko Grobelnik. Extracting Summary Sentences Based on the Document Semantic Graph. Microsoft Research Technical Report MSR-TR-2005-07, 2005.

Jurij Leskovec, Marko Grobelnik, and Natasa Milic-Frayling. 2004. Learning Sub-structures of Document Semantic Graphs for Document Summarization. In Proceedings of the Workshop on Link Analysis and Group Detection (LinkKDD), 2004

Simon Corston-Oliver. 2001. Text compaction for display on very small screens. In Proceedings of the Workshop on Automatic Summarization, NAACL 2001.

Evaluation

Eric Ringger, Robert C. Moore, Eugene Charniak, Lucy Vanderwende, and Hisami Suzuki. 2004. Using the Penn Treebank to Evaluate Non-Treebank Parsers. In Fourth International Conference on Language Resources and Evaluation (LREC'04)

Entailment

Lucy Vanderwende, Arul Menezes, and Rion Snow. 2006. Microsoft Research at RTE-2: Syntactic Contributions in the Entailment Task: an implementation. In Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop, 2006

Rion Snow, Lucy Vanderwende, and Arul Menezes. 2006. Effectively using syntax for recognizing false entailment. In Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)

Knowledge Base Construction / Information Extraction / Text mining

A Kumaran, Ranbeer Makin, Vijay Pattisapu, Shaik Sharif, Gary Kacmarcik, and Lucy Vanderwende. 2006. Automatic Extraction of Synonymy Information. In *the Ontologies in Text Technology Workshop, Osnabruck, Germany*, December 2006

Chris Quirk, Pallavi Choudhury, Michael Gamon, and Lucy Vanderwende. MSR-NLP entry in BioNLP Shared Task 2011. In Proceedings of the BioNLP Shared Task 2011 Workshop.

Spelling Correction

Andi Wu, George Heidorn, Zixin Jiang, and Terence Peng. 2001. Correction of Erroneous Characters in Chinese Sentence. In International Conference on Chinese Computing.

Information Retrieval

Simon H. Corston-Oliver and William B. Dolan. 1999. Less is more: Eliminating index terms from subordinate clauses. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics

Natasa Milic-Frayling, Ralph Sommerer. 2001.*MS Read: Context Sensitive Document Analysis in the WWW Environment*,. SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2001.

Jianfeng Gao and Jian-Yun Nie, 2006. Study of Statistical Models for Query Translation: Finding a Good Unit of Translation. In SIGIR.

Ingrid Zukerman and Eric Horvitz. 2001. Using Machine Learning Techniques to Interpret WH-questions. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, ACL-01.

Intelligent Agents

Simon Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. 2004.Integration of Email and Task Lists. In First Conference on Email and Anti-Spam (CEAS), 2004 Proceedings

Tim Paek and Eric Horvitz. 1999. Uncertainty, Utility, and Misunderstanding: A Decision-theoretic Perspective on Grounding in Conversational Systems. In AAAI Technical Report FS-99-03.

Hua Li, Dou Shen, Benyu Zhang, Zheng Chen, and Qiang Yang. 2006. Adding Semantics to Email Clustering. In Proceedings of the Sixth International Conference on Data Mining (ICDM'06).

Application to Education

Lee Schwartz, Takako Aikawa, and Michel Pahud. 2004. Dynamic Language Learning Tools. Proceedings of the 2004 InSTIL/ICALL Symposium, June 2004.

Takako Aikawa, Lee Schwartz, and Michel Pahud. 2005.NLP Story Maker. In Proceedings of the Second Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics April 21-23, 2005, Poznań, Poland

Sentiment

Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceeding of COLING-04, the 20th International Conference on Computational Linguistics

Authorship identification

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In Proceedings of COLING-2004