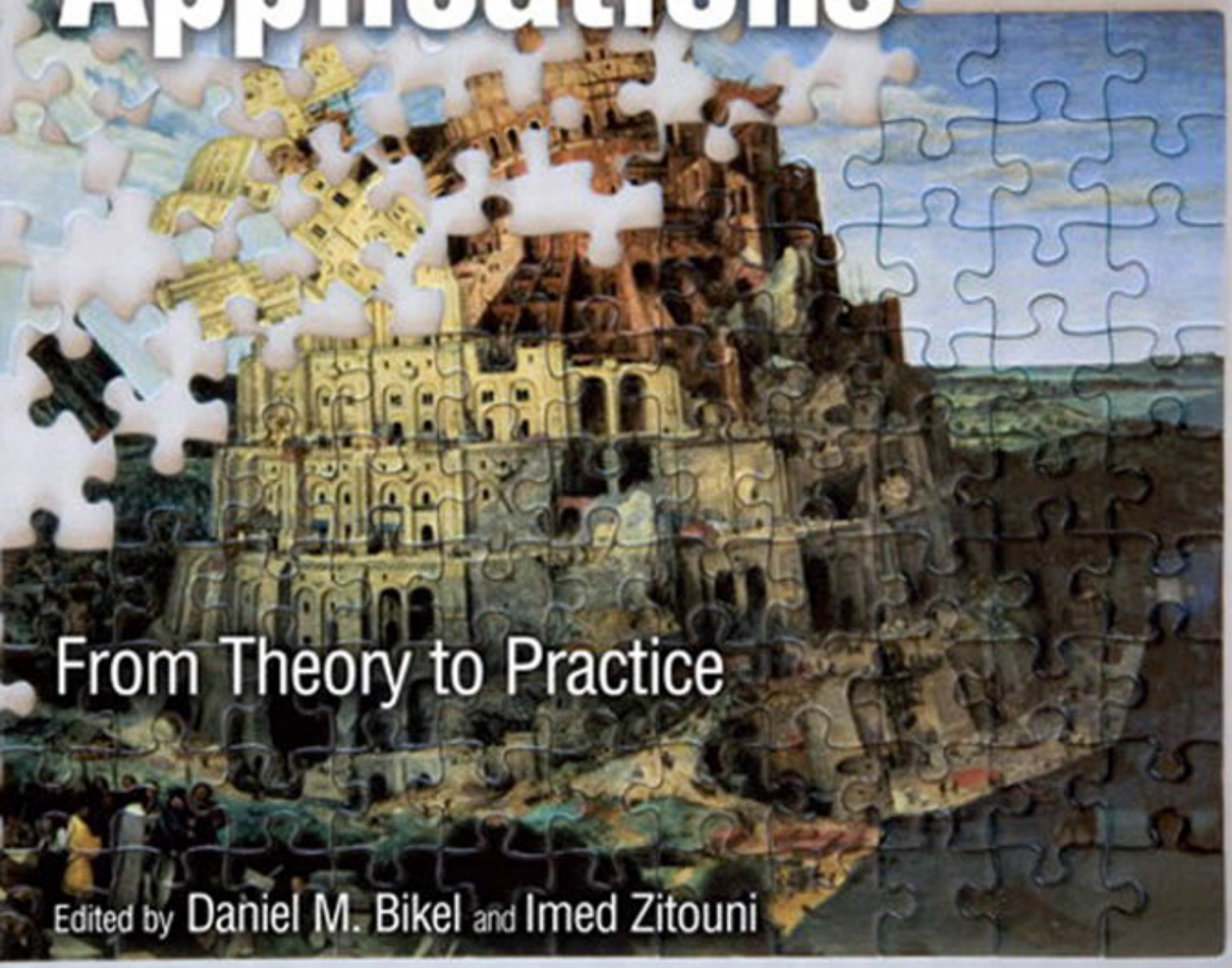
Multilingual Natural Language Processing Applications



From Theory to Practice

Edited by Daniel M. Bikel and Imed Zitouni

Register Your Book

at ibmpressbooks.com/ibmregister

Upon registration, we will send you electronic sample chapters from two of our popular IBM Press books. In addition, you will be automatically entered into a monthly drawing for a free IBM Press book.

Registration also entitles you to:

- Notices and reminders about author appearances, conferences, and online chats with special guests
- Access to supplemental material that may be available
- Advance notice of forthcoming editions
- Related book recommendations
- Information about special contests and promotions throughout the year
- Chapter excerpts and supplements of forthcoming books

Contact us

If you are interested in writing a book or reviewing manuscripts prior to publication, please write to us at:

Editorial Director, IBM Press
c/o Pearson Education
800 East 96th Street
Indianapolis, IN 46240

e-mail: IBMPress@pearsoned.com

Visit us on the Web: ibmpressbooks.com

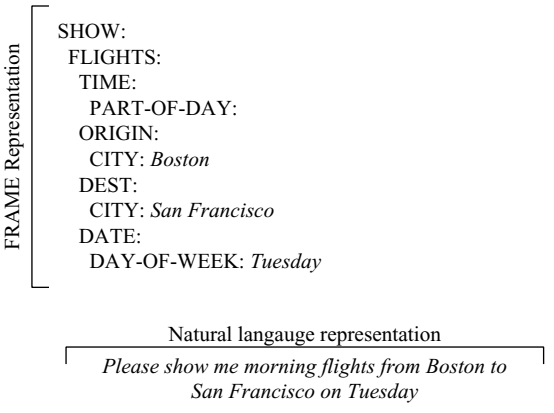


Figure 4–20: A sample user query and its *frame* representation in the ATIS program

GeoQuery

In the domain of U.S. geography, there is a natural language interface (NLI) to a geographic database called Geobase [175], which has about 800 Prolog facts stored in a relational database with geographic information such as population, neighboring states, major rivers, and major cities. Some sample queries and their representations are as follows:

(1) What is the capital of the state with the largest population?
`answer(C, (capital(S, C), largest(P, (state(S), population(S, P))))`

(2) What are the major cities in Kansas?
`answer(C, (major(C), city(C), loc(C, S), equal(S, stateid(kansas))))`

This is the GeoQuery corpus, which has also been translated into Japanese, Spanish, and Turkish.

Robocup: CLang

RoboCup (www.robocup.org) is an international initiative by the artificial intelligence community that uses robotic soccer as its domain. There is a special formal language, CLang, which is used to encode the advice from the team coach, and the behaviors are expressed as if-then rules. Following is an example representation in this domain:

(1) If the ball is in our penalty area, all our players except player 4 should stay in our half.
`((bpos (penalty-area our)) (do (player-except our 4) (pos (half our))))`

4.6.2 Systems

As we can see in these examples, depending on the consuming application, the meaning representation can be a SQL query, a Prolog query, or a domain-specific query representation.

Now we look at the various ways the problem of mapping the natural language to such meaning representation has been tackled.

Rule Based

Some of the semantic parsing systems that performed very well for both the ATIS and Communicator projects were rule-based systems in the sense that they used an interpreter whose semantic grammar was handcrafted to be robust to speech recognition errors. The underlying philosophy was that the traditional syntactic explanation of a sentence is much more complex than the underlying semantic information, so parsing the meaning units in the sentence into semantics proved to be a better approach. Furthermore, especially in dealing with spontaneous speech, the system has to account for ungrammatical instructions, stutters, filled pauses, and so on. Word order therefore becomes less important, which leads to meaning units scattered in the sentences/utterances and not necessarily in the order that would make sense to a syntactic parser. Ward's [176, 177, 178] system, Phoenix, uses recursive transition networks (RTNs) [179] and a handcrafted grammar to extract a hierarchical frame structure, and reevaluates and adjusts the values of these frames with each new piece of information obtained. This system had an error rate of 13.2% for spontaneous speech input with a speech recognition word-error rate of 4.4%, and a 9.3% error for transcript input.

Supervised

Although rule-based techniques are relatively easy to craft in the beginning and serve a good purpose to formulate solutions to various tasks, they have several downsides: (i) they need some effort upfront to create the rules, (ii) the time and specificity required to write rules usually restricts the development to systems that operate in limited domains, (iii) they are hard to maintain and scale up as the problem becomes more complex and more domain independent, and (iv) they tend to be brittle. The alternative is to use statistical models derived from hand-annotated data. However, unless some hand-annotated data is available, statistical models cannot be used to deal with unknown phenomena. During the ATIS evaluations, some data was hand-tagged for semantic information. Schwartz et al. [180] used this as an opportunity to create what was probably the first end-to-end supervised statistical learning system for the ATIS domain. They had four components in their system: (i) semantic parse, (ii) semantic frame, (iii) discourse, and (iv) backend. This system used a supervised learning approach combined with quick training augmentation through a human-in-the-loop corrective approach to generate slightly lower quality but more data for improved supervision. Miller et al. [181] described the algorithm in more detail. Their system achieved an error rate of 14.5% on the entire test set and 9.5% on the subset of sentences that were context independent. Since then, various improvements have been made, such as by He and Young [182].

Continuing on the line of research that is today commonly known as natural language interface for databases (NLIDB), Zelle and Mooney [183] tackled the task of retrieving answers from a Prolog database by converting natural language questions into Prolog queries

in the domain of GeoQuery. They introduced a system called CHILL (Constructive Heuristics Induction for Language Learning), based on the relational learning techniques of inductive logic programming. It uses a shift-reduce parser to map the input sentence into parses expressed as a Prolog program. They preferred a representation closer to formal logic rather than SQL, because once achieved, it can easily be translated into other equivalent representations. They tested the system performance with the rule-based system Geobase that comes with the GeoQuery dataset over a varying number of queries as inputs. It took CHILL roughly 175 training queries to match the performance of Geobase. Additional queries made it surpass Geobase, achieving an accuracy of 84% on novel queries, at times inducing 1,100 lines of Prolog code.

Since then, advances have been made in machine learning and syntactic parsing, and researchers have identified new approaches and refined existing approaches. The SCISSOR (Semantic Composition that Integrates Syntax and Semantics to get Optimal Representation) system, for example, uses a statistical syntactic parser to create a **semantically augmented parse tree** (SAPT) [184, 185]. Training for SCISSOR consists of a (natural language, SAPT, meaning representation) triplet. It uses a standard syntactic parser augmented with semantic tags, then a recursive procedure is used to compositionally construct the meaning representation for each node in the tree given the representations of its children. SCISSOR shows significant performance improvement over earlier approaches. KRISP (Kernel-based Robust Interpretation for Semantic Parsing) [186] uses string kernels and SVMs to improve the underlying learning techniques. WASP (Word Alignment-based Semantic Parsing) [187] takes a radical approach to semantic parsing by using state-of-the-art machine translation techniques to learn a semantic parser. Wong and Mooney treat the meaning representation language as an alternative form of natural language and use GIZA++ to produce an alignment between the natural language and a variation of the meaning representation language. Complete meaning representations are then formed by combining these aligned strings using a synchronous CFG (SCFG) framework. SCISSOR is somewhat more accurate than WASP and KRISP, which can themselves benefit from the information in SAPTs [188]. KRISP, CHILL, and WASP have also learned semantic parsers for Spanish, Turkish, and Japanese with similar accuracies. Yet another approach comes from Zettlemoyer and Collins [189], who trained a structured classifier for natural language interfaces by learning probabilistic combinatory categorial grammar (PCCG) along with a log-linear model that represents the distribution over the syntactic and semantic analysis conditioned on the natural language input.

4.6.3 Software

Not a lot of software programs are available for the older, more rule-based systems, but the following are available for download.

- **WASP** [http://www.cs.utexas.edu/~ml/wasp/]
- **KRISPER** [http://www.cs.utexas.edu/~ml/krisp/]
- **CHILL** [http://www.cs.utexas.edu/~ml/chill.html]

4.7 Summary

In this chapter we looked at the problem of semantic parsing through various lenses. There is no silver bullet to the problem of meaning representation and language understanding, so over the years, researchers have come up with tasks that either solve parts of the bigger problem in a more domain-independent fashion or solve the complete problem but for a very restricted domain. The first case, shallow semantic interpretation, deals separately with the four main aspects of language: structural ambiguity (which is syntactic in nature and is the subject of a separate chapter), word sense, entity and event recognition, and predicate-argument structure recognition. The latter three are components of what has widely come to be known as shallow semantic parsing. As we have seen, syntax plays a very important role in this process, and cannot be considered completely divorced from semantics. The second, deep parsing, or semantic parsing, comprises taking natural language input and transforming it into a meaning representation, which tends to be task specific and something that the end application can unambiguously execute.

We learned that developments on various fronts have been made in all these methods. In the early era of the field, few hand-labeled corpora and few well-developed learning techniques were avaliable. Even now, in the case of resource-poor languages, there is not much data to train sophisticated learning algorithms. In these cases, researchers resort to encoding the domain information in a rule-based system, which is usually domain specific. For languages for which there is enough human-annotated data available, more statistical approaches became predominant. Given the sparseness of data even when there is sufficient annotation (any amount of realistic human annotation would never be enough to learn all the various nuances of language), researchers have resorted to semisupervised or unsupervised methods, the latter being usually much less accurate than supervised or rule-based methods.

4.7.1 Word Sense Disambiguation

Word sense disambiguation is an integral part of language understanding. In information retrieval, speech understanding, and applications restricted to certain domains, it has not been very important, because of limited sense usability or implicit disambiguation. However, for applications that deal with a deeper understanding of text, sense disambiguation may be critical. Research in this area started with senses that were defined in dictionaries, because they were the primary resource in the beginning. Lesk's algorithm is generally recognized as the first dictionary-based word sense algorithm, where sense disambiguation is performed using the overlap between the context in which a word appears in the discourse and its dictionary gloss. The creation of *Roget's Thesaurus* led to more English-specific algorithms to classify words into the categories defined in it. The notion of one sense per discourse led to an important semisupervised algorithm: the Yarowsky algorithm. With the advent of a much richer lexicon like WordNet and corpora annotated with senses defined in it (SEMCOR)—interestingly, in parallel with the advances in machine learning—most of the research community shifted to using them as the standard until later studies showed that the granularity of senses in WordNet might be too fine. If humans cannot agree on sense distinctions to a certain degree, then machines should not be expected to either. This led to the folding of word senses in WordNet into coarser units that are more amenable to

human agreement and at the same time provide a better means of achieving high-accuracy automatic disambiguation. WordNet has continued to be an important resource that has significantly improved the field, and it is still used in state-of-the-art disambiguation systems.

In a separate vein, with the growth of the Internet, the availability of resources such as Wikipedia, which served as a surrogate annotation resource, exploiting Internet resources became one of the mainstream pursuits. An increasing number of areas in language understanding are making use of this resource in novel ways. Active learning is another direction that is still probably more an art than a science but has been quite useful for amassing annotations for words that are either rare (low-frequency), high sense perplexity (many senses), or do not have enough annotation for some reason or other, including, but not limited to, low-resource languages [190]. In languages where there was no hand-annotated data available, various unsupervised approaches were developed, some of which exploited the differing sense granularities and instantiations across parallel corpora.

4.7.2 Predicate-Argument Structure

Unlike word sense disambiguation, there were far fewer rule-based systems that tagged thematic roles in text. With the advent of corpora such as FrameNet and PropBank labeled with a predicate-argument structure, there was a giant wave of research focused on building systems to tag these structures in text, primarily for verb and noun predicates. Many new features were introduced in various syntactic frameworks, some of them not even conducting a full syntactic analysis and resorting only to the base phrase chunks. It turns out that, at least for the genres that have treebanks, the contribution of a syntactic parser is invaluable. When lexicalization is the only way in which a syntactic representation is informed by semantics, a semantic role labeler tends to make errors that could be avoided using a more bottom-up approach. Also, using richer features in the first pass can be prohibitive, so a combination of producing n-best hypotheses and reranking them based on a more global feature set is the approach that generally performs better. Furthermore, a combination of a top-down and bottom-up approach by combining the information from various syntactic and nonsyntactic views improves performance as well. One big bottleneck at present is that performance on genres of text that exhibit even somewhat different syntactic styles, word usage, or entity and event structures tends to be much worse than if you have matched training and test corpora. We are at a point where structural information has been utilized to a significant degree to the benefit of semantic analysis, but the lexical- and sense-level generalization is significantly lacking, thereby making the existing approaches much less robust across genres or domains of text. We also saw that the fundamental techniques developed for English—which happens to be the language for which the hand-tagged corpora were first created—translate very well to other languages. Of course, each new language has its own idiosyncrasies that lead to the identification of new features. These new features may in turn improve the original English system. Many annotation efforts are underway across the world, and we have much more to learn.

4.7.3 Meaning Representation

Finally, we looked at meaning representation. This is a much less researched topic and especially so across languages. Meaning representation is the process of converting natural

language input into a format that is unambiguous and easily understandable by a machine or end application, which can take actions based on the input. So far, there is no one universally accepted representation, so these systems and their representations tend to be domain specific.

New research programs are stretching the possibilities of existing techniques and creating novel ones that will eventually let us bring these pieces together into a richer, deeper representation that would also be independent of domain.

Bibliography

[1] N. Chomsky, *Syntactic Structures*. The Hague: Mouton, 1957.

[2] J. Katz and J. Fodor, "The structure of a semantic theory," *Language*, vol. 39, pp. 170–210, 1963.

[3] V. H. Yngve, "Syntax and the problem of multiple meaning," in *Machine Translation of Languages* (W. N. Locke and A. D. Booth, eds.), pp. 208–226, Cambridge, MA: MIT Press, 1955.

[4] N. Ide and J. Véronis, "Introduction to the special issue on word sense disambiguation: The state of the art," *Computational Linguistics*, vol. 24, no. 1, pp. 2–40, 1998.

[5] E. Agirre and P. Edmonds, eds., *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer, 2006.

[6] M. Palmer, H. Dang, and C. Fellbaum, "Making coarse-grained and fine-grained sense distinctions, both manually and automatically," *Natural Language Engineering Journal*, vol. 13, no. 2, pp. 137–163, 2007.

[7] P. Resnik and D. Yarowsky, "Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation," *Journal of Natural Language Engineering*, vol. 5, no. 2, pp. 113–133, 1999.

[8] R. Krovetz and W. B. Croft, "Lexical ambiguity and information retrieval," *ACM Transactions on Information Systems*, vol. 10, no. 2, pp. 115–141, 1992.

[9] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 3, pp. 400–401, 1987.

[10] L. Bahl, F. Jelinek, and R. Mercer, "A maximum likelihood approach to continuous speech recognition," *PAMI—IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 179–190, 1983.

[11] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[12] W. Gale, K. W. Church, and D. Yarowsky, "Estimating upper and lower bounds on the performance of word-sense disambiguation programs," in *Proceedings of the*