Microsoft Research Tackles the Universal Translator:

A Breakthrough in Accurate and Automatic Translation

By Don Barker

Introduction

By the close of 2002, Bill Dolan, manager of Natural Language Processing (NLP) Group at Microsoft Research (MSR), expects the software titan to deploy a Web-based system capable of accurately and automatically translating the entire Microsoft Product Support Services (PSS) Knowledge Base from English into Spanish, providing real-time responses to Spanish queries. The *PSS Knowledge Base* is an enormous collection of information used to identify and solve problems with Microsoft software.

Accurately and automatically converting this sizable knowledge base from English to Spanish, without human editing, represents an incredible leap forward in the world of computational linguistics. This breakthrough, made possible by the fruition of a two-decade plus research effort, produced an NLP system known internally at Microsoft as "NLPWin."

The system has already been successfully beta tested. The NLP Group is now working on English to Japanese, German, French, and Chinese versions of the NLPWin executable (program). These systems will likely save the company millions of dollars and that is just the beginning.

The NLP Group is exploring how its Machine Translation (MT) technology can help product divisions within Microsoft. "Demand for the technology is far outpacing the capacity of our 30 member research group to satisfy requests," says Dolan. One Microsoft product team especially interested in the technology is the Productivity Tools Group, which has the daunting task of product "localization."

Localization, that lengthy process of preparing software for a foreign market, typically involves translating the text embedded in the software itself (e.g. menus, message dialog boxes) as well as the associated user manuals and other help text into the native language. As you can imagine, this is a considerable undertaking for a product like Office XP. The NLP Group's MT technology has the potential for making this time-consuming, complex, and expensive procedure fast, simple, and inexpensive.

They also see their innovation eventually being packaged for use by other large corporations in need of translating sizable bodies of documents quickly and cheaply. Some in the NLP Group envision a time when a "mega-translator," based on their technology, will allow Internet users to converse in unrestricted domains instantaneously. Unleashing this type of communication power for public use could open a entirely new world of global interactions.

A Little History

To better appreciate and understand the implications of Microsoft's Machine Translation breakthrough, it is helpful to briefly examine the evolution of the field. The quest for accurate, automatic, on the fly MT has been the Holy Grail of leading computational linguistics and AI researchers for over fifty years. The effort began when Warren Weaver, then director of the Rockefeller Foundation, wrote a 1949 memorandum to 200 top scientists, suggesting that computers could be programmed to translate language mathematically, without actually "understanding" the meaning of words. This seminal 12-page memo literally launched the field of MT.

Within in a couple of years, MT efforts were underway at UCLA, the National Bureau of Standards, the University of Washington, the Rand Corporation, and MIT. In 1953, a Georgetown University team worked with IBM to actually create the first working MT program, which translated Russian into English — the language choices were no doubt inspired by the Cold War atmosphere of that era. On January 7, 1954, the Georgetown team unveiled the MT program publicly at IBM's Technical Computing Bureau in New York. Despite the fact that it was limited to just 250 words, 6 grammar rules, and 49 handpicked sentences, the idea of MT caught fire in the press.

When the Soviets successfully placed the first satellite, "Sputnik," into earth orbit in 1957, it took America by surprise. Behind the scenes, the news was more distressing because technical details of the satellite had actually appeared months before its launch in a Soviet hobby magazine. However, it had gone unnoticed because American intelligence did not have the means to quickly translate Russian to English.

While not nearly as high profile or sexy as the race to the moon in the 60s, the government also set out to achieve automatic translation of Russian into English, dumping millions into academic and industrial MT research. Agencies, such as the CIA, relished the thought of having immediate access to thousands of Soviet papers and publications, with an eye on the huge advantage it would give them in counter-intelligence.

Unfortunately, almost a decade and 20 million dollars later, results were not meeting these overly optimistic expectations. In 1966, the Automated Language Processing Advisory Committee (ALPAC) issued a highly critical report, citing the lack of significant progress, which effectively halted government spending on MT.

Star Trek – Universal Translator

In part, one can attribute the apparent failure of MT research to the unrealistic expectations set in this field's early days. The public's first general introduction to the concept of MT came from the classic 1960s TV series Star Trek, where the crew of the starship Enterprise used a device called the "Universal Translator" to communicate with alien races across the galaxy.

With little more than a few snippets of dialogue from a newly encountered race of sentient beings, the Universal Translator deduced the meaning of their languages entire lexicon and flawlessly, in near real time, translated speech. In retro-spect, not only was this unrealistic for the times, but a downright impossible goal.

Fully automatic, high quality text-to-text Machine Translation across vastly different knowledge domains is challeng-ing. However, throw in a scarcity of training data and speech-enabled front and back ends, and the ideal symbolized by the Universal Translator becomes unachievable even with today's best tech-nology. Due to overly optimistic expect-ations and a subsequent collapse of government funding, research into MT survived in only a few institutions that could afford going it alone, such as IBM and strangely enough -- the Mormon Church.

The Mormon Connection

In the late 1970s, the Church of Latter Day Saints undertook a massive MT project in hopes of making it relatively easy to translate their religious literature into different languages. A key figure in that effort was Steve Richardson, first an undergraduate and then graduate student at Brigham Young University in Provo, Utah, who used his computer science and linguistic education to further the Mormon MT effort.

Upon completion of his bachelor's degree in 1977, Richardson worked full-time for the Mormon MT project until he completed his master's degree in 1980. At that point, after five years, they canceled the undertaking. Although not successful at producing cost-effective MT because of the high cost of computing power on the IBM mainframes, the project inspired a number of MT start-ups in the Utah area, the descendents of which continue in operation today. With a growing family to support, Richardson took a job as an associate program-mer with an IBM product group in Endicott, New York.

The IBM Connection

In 1983, Richardson contact-ed a group at IBM's famous T. J. Watson Research Center, dedicated to pushing the limits of natural language processing. Richardson met George Heidorn and Karen Jensen on an incredibly snowy day in mid-February. "I remember my first meeting with Karen and George clearly, on February 11, 1983, because it was snowing so hard that the Watson Research Center had to close," says Richardson.

Heidorn was the manager of the Natural Language Processing group at IBM Watson and Jensen, a leading authority in English grammar, his close colleague. Heidorn, Jensen, and Richardson formed a powerful trio of talent that weathered many technical and corporate storms to eventually build their shared dream — one of the largest and most successful Natural Language Processing (NLP) Projects in the world.

In the 1980s, government and industry funding was again flowing for MT research and development. The launch of the Japanese Fifth Generation Project, aimed at building an intelligent computer within 10 years, was the equivalent of another "Sputnik-scare," spurring the U.S. government to again open its purse strings for MT research. This funding launched large-scale efforts, such as CYC (short for encyclopedia), to create software capable of understanding natural language via common sense reasoning.

Realizing the need to demonstrate the practical application of their NLP Project, the IBM trio transferred to Big Blue's development side, with

hopes of including a grammar checker in a software suite tentatively dubbed "OfficeVision," intended to compete directly with Microsoft's highly successful Office.

The Beginnings at Microsoft Research

However, IBM OfficeVision effort buckled under the company's stifling bureaucracy, and its collapse threatened to end Heidorn, Jensen, and Richardson's NLP Project. In a bold move, the trio jumped ship and moved the project to Microsoft's fledgling research lab in 1991, becoming the first researchers to join Microsoft Research (MSR) and the first to leave the prestigious halls of IBM Watson for the relatively small software company, mainly known for its DOS program.

To add insult to injury, this move took place at time when IBM and Microsoft were going through a very public "divorce" over the OS/2 personal computer operating system. Microsoft had decided to abandon the joint OS/2 development effort in favor a Windows-only strategy. Microsoft's decision was widely viewed within IBM as nothing less than base treachery, and losing three top researchers to Microsoft during this period made matters all the more infuriating.

Hence, Heidorn, Jensen, and Richardson's transition to MSR was fraught with nasty episodes, such as IBM locking them out of their offices. The rights to the group's intellectual property were a real point of contention between the two companies. Fortunately for the trio, they had placed most of the work on the NLP Project in the public domain in various scientific journals and publications. Nonetheless, the trio had to coauthor a book to document this fact before IBM relinquished its claims to the concepts.

Once at MSR, the trio immediately began recruiting theoretical and computational linguists, starting with the team members they had originally assembled at IBM. Bill Dolan, Lucy Vanderwende, and Joseph Pentheroudakis, the first three recruits, were deemed critical in moving the NLP Project along — and they possessed sufficient computer skills to pass muster with Microsoft's development managers. Dolan and Vanderwende had worked with the technology at IBM, while Pentheroudakis had been at Executive Communications Systems (ECS), a leading developer of natural language software.

The NLPWin System

The NLP Group at MSR, growing steadily in number, strove to make their NLP Project's conceptual modules a reality in the NLPWin system. This systems conceptual framework consists of the components shown in Figure 1, with each module in this series designed to successively abstract the structure and meaning of the words and sentences within the language.

The NLP Group's bottom up approach to natural language processing contrasts sharply with the approach typically taken by researchers in the Artificial Intelligence (AI) field. AI scientists primarily focus on creating a machine that reasons similar to a human and on reach this ability to thinking, assume it is a relatively trivial task to generate a natural language dialogue with humans. However, after forty plus years of AI research, this top down approach has shown little success. The mapping between the abstract concepts found in machine reasoning and the highly rich, complex nature of natural language is much more difficult than first imagined.

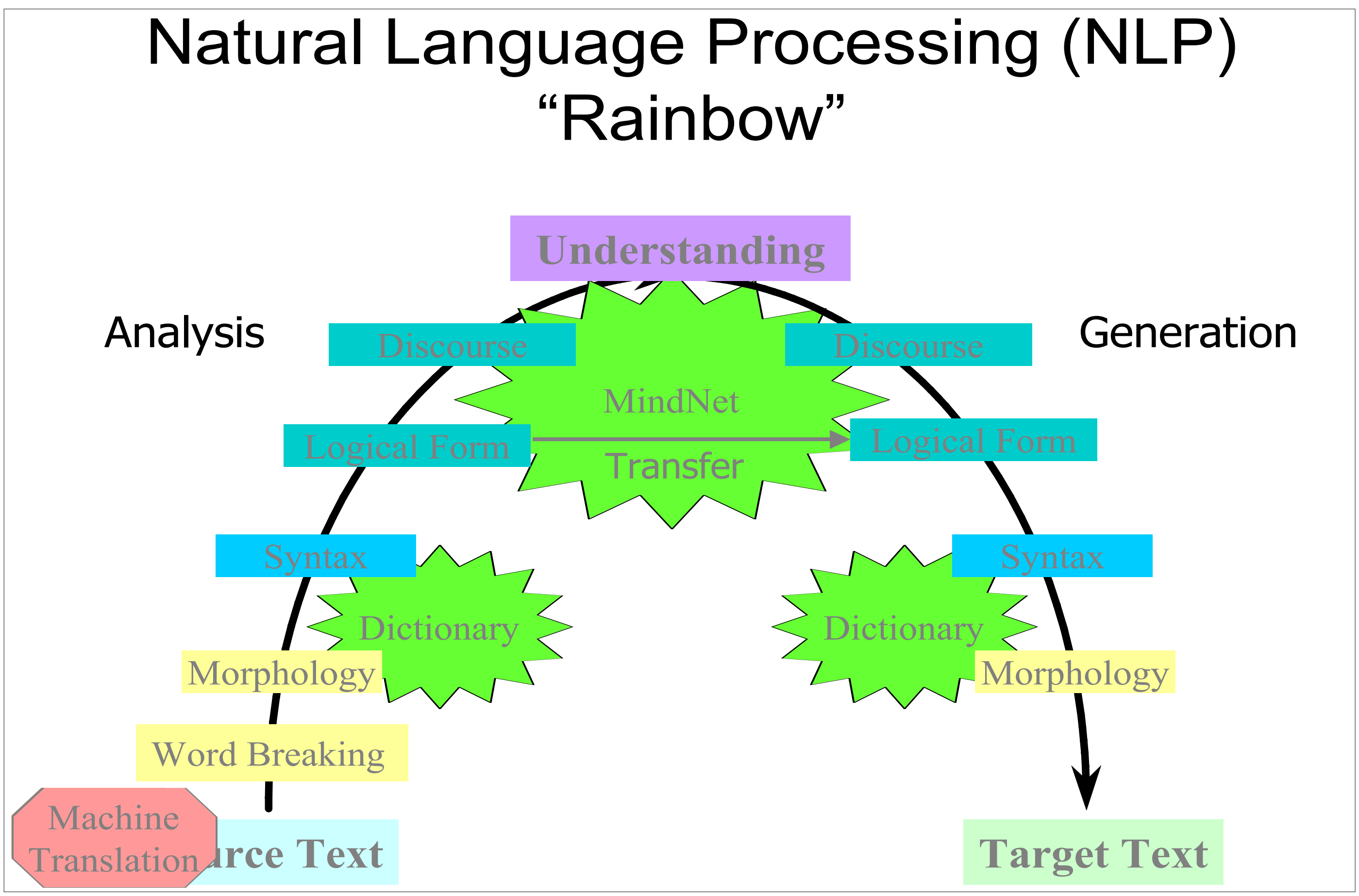


Figure 1: NLPWin system conceptual framework and components.

"The goal of NLPWin is to enable the machine to produce an internal representation that corresponds to what we understand in our minds when we hear natural language. This is the key — the understanding of natural language leads to intelligence. I do not think humans become intelligent just through natural language. I think as we are born we take in all kinds of sensor input. We have emotions that are just native to react with, we learn by being immersed in this environment, language comes along and we put the symbols on these experiences. Machines are not like that, if they are going to become intelligent, it is going to have to be some other way. Therefore, our way is through experience and their way is through symbol manipulation. We put symbols on our experience and machines are going to have to learn to put experiences on the symbols," says Karen Jensen, former manager of the NLP Group.

This bottom-up vision for building intelligent machines flies in the face of large-scale top-down AI efforts such as the 18-year-old CYC project pioneered by AI legend Doug Lenat. A second major area of difference between NLPWin and CYC is in self-training. The NLP Group strongly believes that CYC's handcrafting is counterproductive. Every time CYC encounters a new lexicon, it requires more hand coding to surgically implant the new knowledge, slowing development and possibly creating conflicting information. Instead, NLPWin automatically assimilates the meaning of words from the text.

Assimilating the Meaning of Words from the Text

This process involves a series of successive stages, beginning with a very rudimentary analysis of how words connect together to form grammatically correct sentences. It then explores the deeper structures in the language hoping to attach meanings to the words and sentences in the context of the world. As shown in Figure 1, the systems first component breaks or parses words, arranging them in a tree-like structure.

The next component, Morphology, identifies the various forms of a word. For example, the root word *jump* has a variety of variations, or morphs, such as *jumping*, *jumped*, and *jumps*. By storing just the root word *jump*, and retaining the capacity to recognize the other morphs of the word, the system saves approximately one half the space it otherwise requires to store all variations of the English words. The savings is even greater for other languages, such as Spanish, Arabic and Japanese, where the savings can run as much as three to four times.

Microsoft Natural Language Dictionary (MIND) Component

Joseph Pentheroudakis designed the Morphology component and the Microsoft Natural Language Dictionary (MIND), which was originally built using

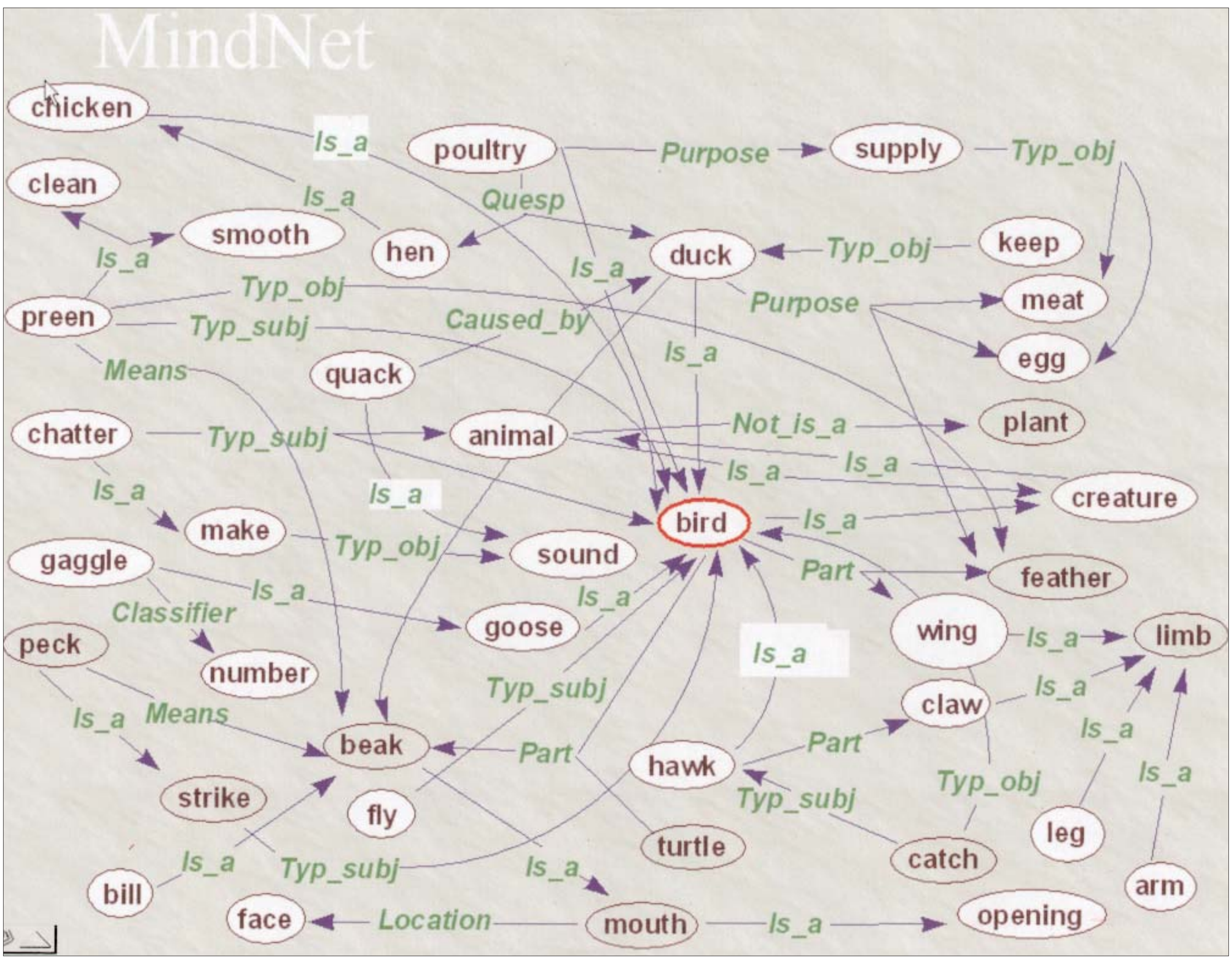


Figure 2: The conceptual view of how words interlock in MindNet.

two different machine-readable dictionaries, the Longman Dictionary of Contemporary English and the American Heritage Dictionary. Although NLPWin uses dictionaries to train itself, "the parser has not been specifically tuned to process dictionary definitions. All enhancements to the parser are geared to handle the immense variety of general text, of which dictionary definitions are simply a modest subset," says Pentheroudakis.

G Component

George Heidorn spearheaded the development of a programming language called "G" (—short for "Gamma", and also for "Grammar" or "George" J). G has a lot in common with the AI language LISP, except that it includes specialized structures for representing the linguistic relationships. G, along with MIND, enabled the NLP Group to transform their conceptual dreams of an NLP system into the reality of a working program, eventually known as NLPWin.

Microsoft English Grammar (MEG) Component

Karen Jensen, a leading authority on English grammar, accomplished the awesome task of creating a comprehensive set of English grammatical rules, using the G language. These rules, called the Microsoft English Grammar (MEG), form the basis of the NLPWin Sketch component. The Sketch module parses text to produce syntactic structures, which are passed to the next component in the system. "The beauty of NLPWin is that any ambiguity is retained and passed up to the next level for resolution there or beyond," says Jensen.

Portrait Component

Lucy Vanderwende oversaw the construction of NLPWin's next stage, Portrait, which uses semantic information automatically extracted from the definitions and example sentences in MIND, to determine correct phrasal attachment during parsing. In other words, the Sketch component does not attach prepositional phrases, but the Portrait component does.

Logical Form Component

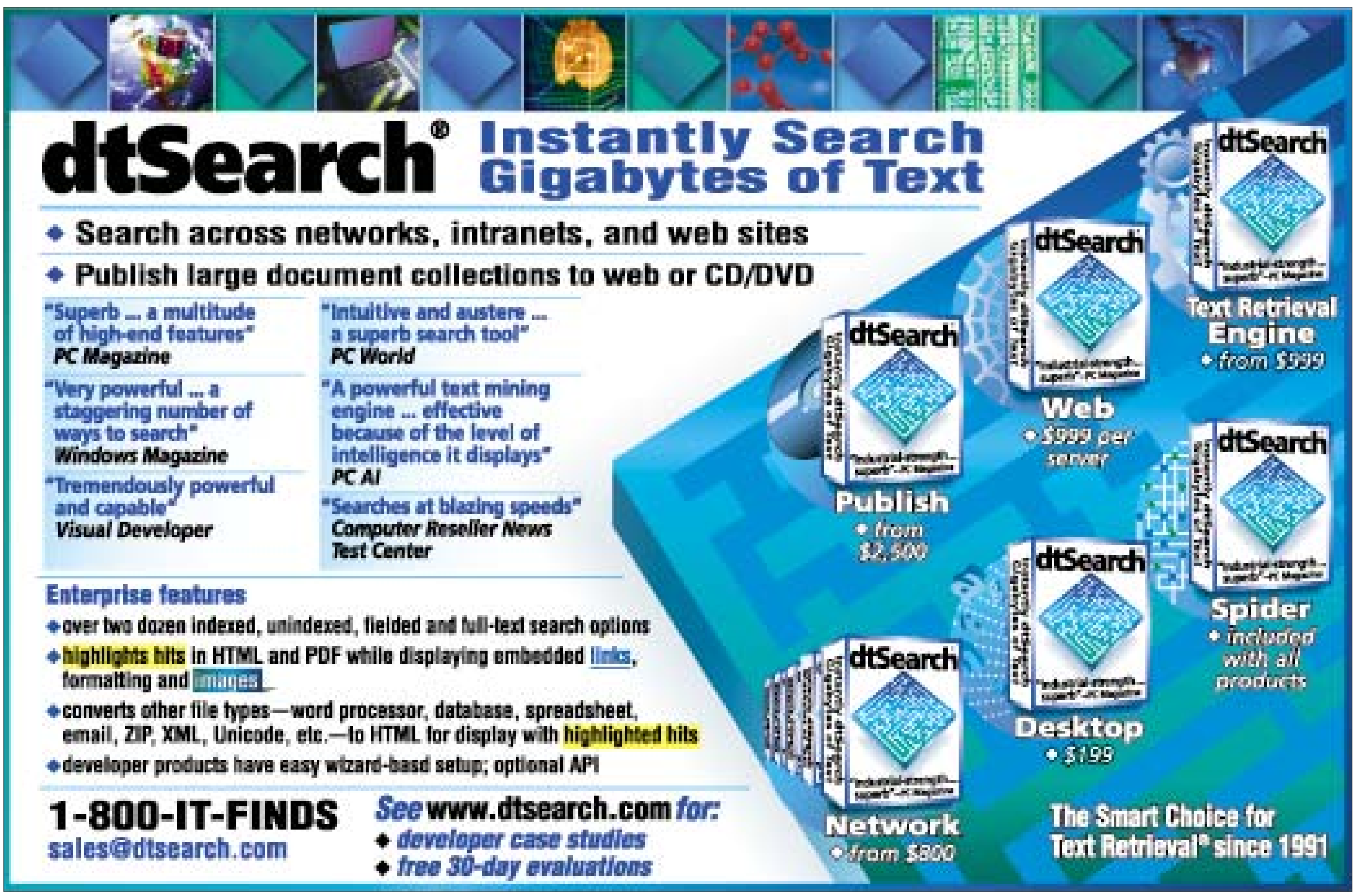
Vanderwende also played a significant role in the development of the Logical Form component. This module encodes the abstract relations between the concepts in a sentence. "Many of these relationships can be captured using a small set of semantic relationships between a head word and its modifiers," says Vanderwende.

Perhaps the single biggest challenge in developing NLPWin was creating the method for storing the mapping of the complex and abstract relationships among words. Although a group effort, Bill Dolan originated the conceptual framework for a semantic network. It had to be capable of representing the inter-linking relationships between the logical forms (grammatical relationships) among words parsed from machine-readable dictionaries and other sources.

Mindnet

Heidorn and Richardson lead the way in turning this theoretical structure into a working code base. The auto-construction of semantic nets was not a new idea in the early 1990s. However, building a program that self trained from a variety of language sources and retained the ambiguity in natural language, critical for discovering the meaning of words, was a radical concept. After years of experimentation, and a number of breakthroughs, the NLP Group finally developed the means to auto-construct a semantic net capable of accomplishing both requirements and called it MindNet.

Figure 2 illustrates the conceptual view of how words interlock in MindNet.



For example, the word *bird* maps to *Hawk* through the *is_a* relationship. *Duck* also interlocks with *bird* by the same is_a relationship. By sliding along these relationships, NLPWin uses the knowledge stored in MindNet to identify the meaning of words in relations to other words.

Discourse Component

The Discourse module, pioneered by Simon Corston-Oliver, takes the data passed up from previous components and summarizes it. For instance, it can summarize the essence of a book, similar to Cliff Notes, presenting the key points of the book.

Meaning Representation Component

At the top of the NLP arch, the Meaning Representation component represents the Holy Grail of computational linguists, true language understanding. Once in this state, NLPWin has finished the increasingly abstract parsing of the original text and it stores the information in MindNet, it is possible to reverse the entire process to produce meaningful responses.

In other words, the Generation component converts the abstract, or logical, forms taken directly from NL Text back into NL Text. By first dissecting and digesting text fed into it and then synthesizing meaningful responses enables the system to engage humans in conversation (dialogue). While many of the previous attempts at this type of system have focused on narrow vocabularies, the NLP Group's ambition is to enable broad coverage of entire natural languages, such as English, Spanish, Japanese, etc.

Applying NLPWin to Machine Translation

Although the research linguists at Microsoft have made groundbreaking strides in developing the initial components of NLPWin (with the Word grammar checker perhaps the most notable milestone), teaching computers to actually understand language remains a distant goal. Given that the language Generation module appears to depend on the Meaning Representation component, the successive and cumulative nature of NLPWin implies that language translation remains beyond the current capabilities of the system.

Fortunately for the field of Machine Translation (MT), the NLP Group has found a method to short-circuit the process. Once it reaches the Logical Form stage, these highly abstract constructs stored in MindNet it is possible to match or map to their counterparts in another language. Thus, the system could perform MT without the machine truly understanding the meaning of the words.

The creation of the NLPWin Machine Translation system takes place in two stages: training and runtime.

Training

Figure 3 presents an overview of the MT training process. The system begins with a pair of equivalent sample sentences

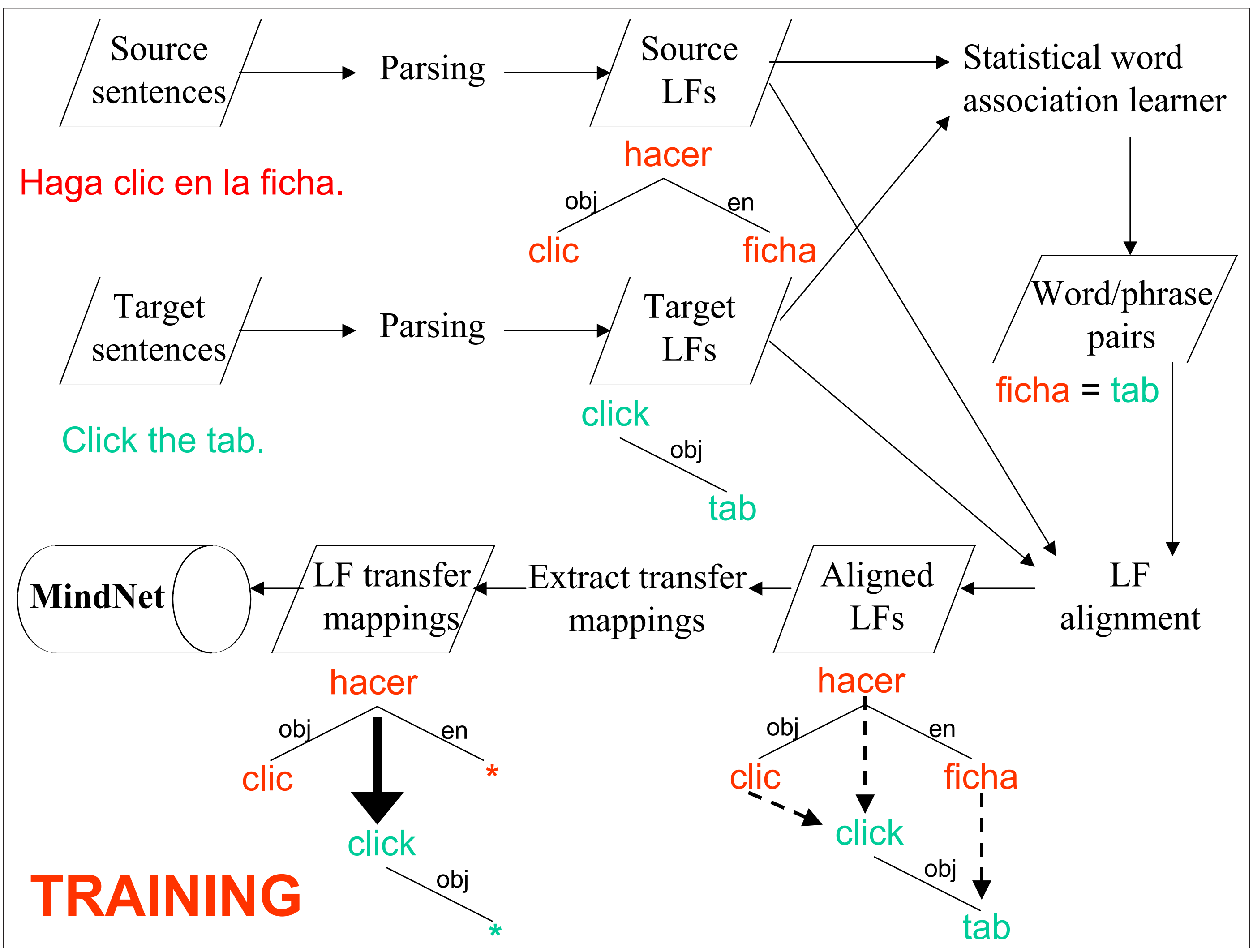


Figure 3: An overview of the MT training process.

from a database. In this example, the Spanish source sentence “Haga clic en la ficha” and the matching English Target sentence “Click the tab” are parsed, using the Morphology, Sketch, Portrait, and Logical Form components, into their respective source and target Logical Forms (LFs). Both LFs undergo statistical processing to identify word associations (e.g., “ficha” and “tab”) and “alignment” of their structures.

“This is done for 350,000 sentence pairs in English and Spanish, applying both heuristics (rules) and statistics to find bits of structural alignment across the language boundary. Most of the MT work has gone into the alignment phase, figuring out which bits across that language boundary should align up and what context you need to save,” says Dolan.

Rules help the system learn the appropriate context, narrowing the search. A probability is attached to each correspondence, or “mapping”, for use during runtime. Finally, these transfer mappings are stored in the MindNet repository.

“We thought it would take us a lot longer to make progress on machine translation. It has come together pretty fast,” says Dolan. To speed development, the research arm of Microsoft took a page from the product side by creating nightly NLPWin builds to make available feedback on progress.

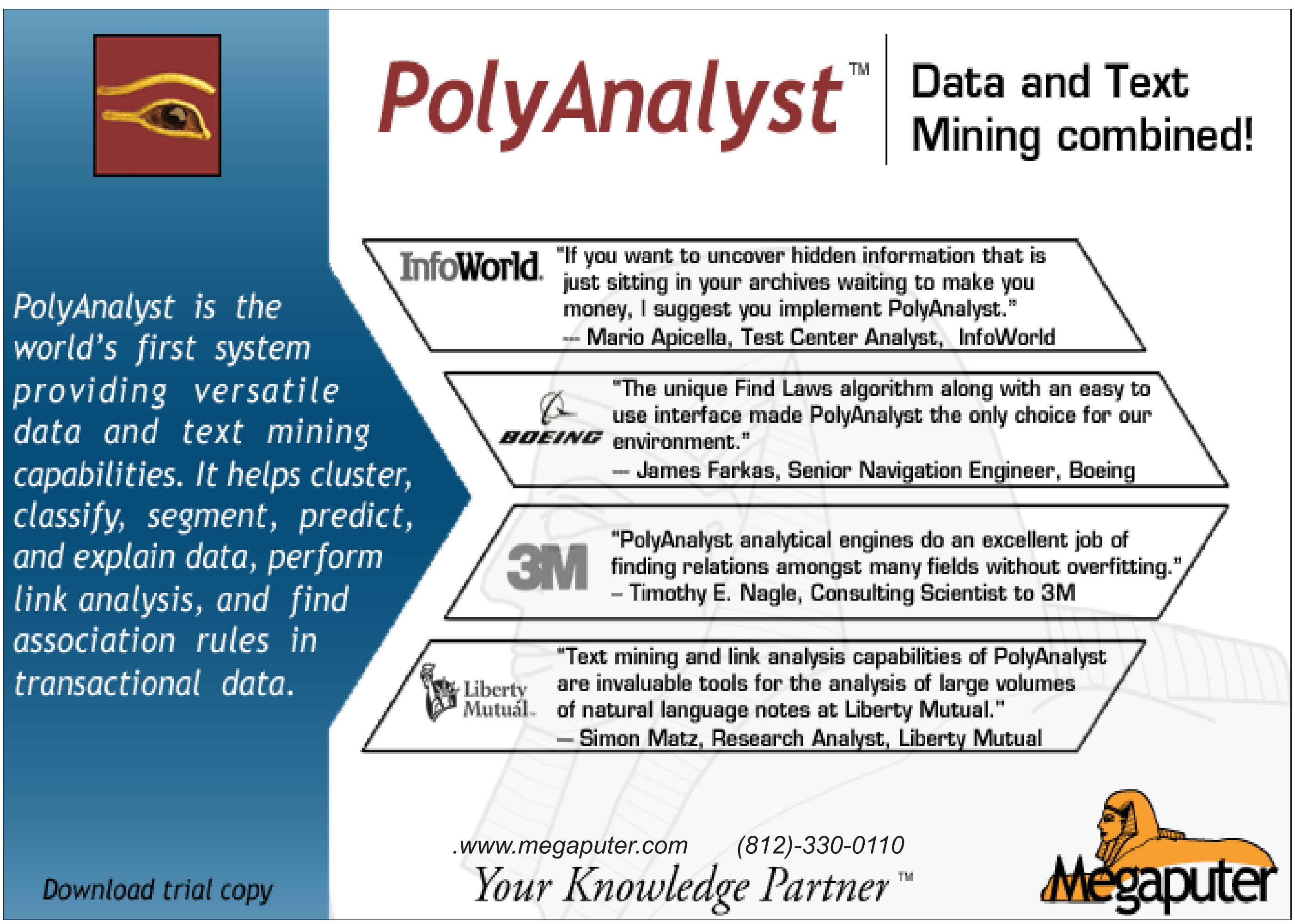
Helping speed this process along, the researchers “…use a huge cluster of 30 computers to retrain the system and rerun a regression test every night,” says Richardson. This produces a new NLPWin build each day, as well as a newly updated version of MindNet. Consequently, the NLP Group sees the impact of the previous day’s work on the MT system’s effectiveness, making it simpler to recognize positive and negative changes to the code, and fixing the latter.

Another longtime obstacle to progress in natural language processing is the lack of an objective means to accurately measure advancements. The typical metric, having humans judge the accuracy of a machine translation, makes the process inherently subjective. The NLP Group developed a more objective testing metric, which compares how close the MT system comes to matching an ideal sentence translation. By minimizes the role human judgment plays in determining MT improvements, this approach is a more quantifiable process, as has been the case in speech recognition for decades.

Runtime

Figure 4 illustrates how the MT system works during runtime. In this example, a Spanish source sentence is parsed by NLPWin into its source LF. The next stage, MindMeld, refers to a highly sophisticated process that has consumed the NLP Group’s research efforts since 1997. “MindMelding takes a sentence and matches it to the closet conceptual relationship in a MindNet,” says Dolan.

This is essentially a graph matching process, which takes an input sentence LF and attempts to match it against one or more subgraphs in MindNet. For instance, if the Spanish source LF is uncomplicated, it might exactly match an English target LF in MindNet. Typically, the match requires paraphrase identification. This involves sliding



around on the lexical similarity dimension to locate a match (e.g., "canine" against "dog"). Syntactic paraphrasing may also come into play (e.g., matching "Jupiter has 18 moons" to "Jupiter's 18 moons"). Often both are required ("How many moons does Jupiter have?" vs. "Jupiter's 18 satellites").

"MindMelding relies on MindNet's path-finding and lexical similarity routines. Briefly, paths between the least frequent word in the input graph and other words directly connect to it are identified. Along these paths, typically, are words that are found to be similar to one of the endpoints (e.g. looking for paths between 'car' and 'top' might provide paths linked through 'vehicle' or 'hood'). These newly-identified words, which aren't simply similar in meaning to the original words but, crucially, similar in this particular lexical context, can now be used for matching if no structures with the original words can be found. This process is iterated, so that a number of contextually-similar words can be identified," says Dolan

Although the MindMelding algorithm relies on the type of graph matching that is intractable (impossible) for the worst cases, the wide variety of context and linguistic heuristics that the MT system brings to bear on the matching problem prevents worst case scenarios from occurring. Nonetheless, carrying out the match efficiently is still a highly complex challenge.

"We take the Logical Form and try to find pieces that match in the stored database mapping [of MindNet] and follow those to the corresponding link on the English side. Grab all those pieces and sort of Frankenstein monster-like put them together into a Linked Logical Form. Right now we are working on using language modeling techniques to smooth out any differences that make that stitched together Logical Form look non-native… using statistical techniques, we smooth out any wrinkles that don't look like what an English Logical Form should look like," says Dolan.+1

Once MindMeld has worked its magic, the corresponding pieces of target LFs are stitched together to form an English target LF, which is handed off to the Generation module. "Provided we've done a good job of assembling a LF, the NLPWin's generation component reliably maps that LF into a well-formed target-language sentence," says Dolan. In the example shown in Figure 4, the English string "Click the highlighted sample text" is generated from the original Spanish input "Haga clic en el texto de muestra resaltado."

At runtime, NLP Group's MT system translates all the English text in the Microsoft Product Support Services Knowledge Base (KB) into Spanish, allowing users to search the converted KB using Spanish queries. "As articles are added or updated in English (which happens a couple thousand times a week), they will be immediately (re-)translated and posted to the Spanish KB. Occasionally, as the MT system improves, the entire KB will be retranslated using newer versions of the MT system. This will happen incrementally, so users should not experience any down time," notes Richardson. To date, internal Microsoft studies indicate a high level of satisfaction

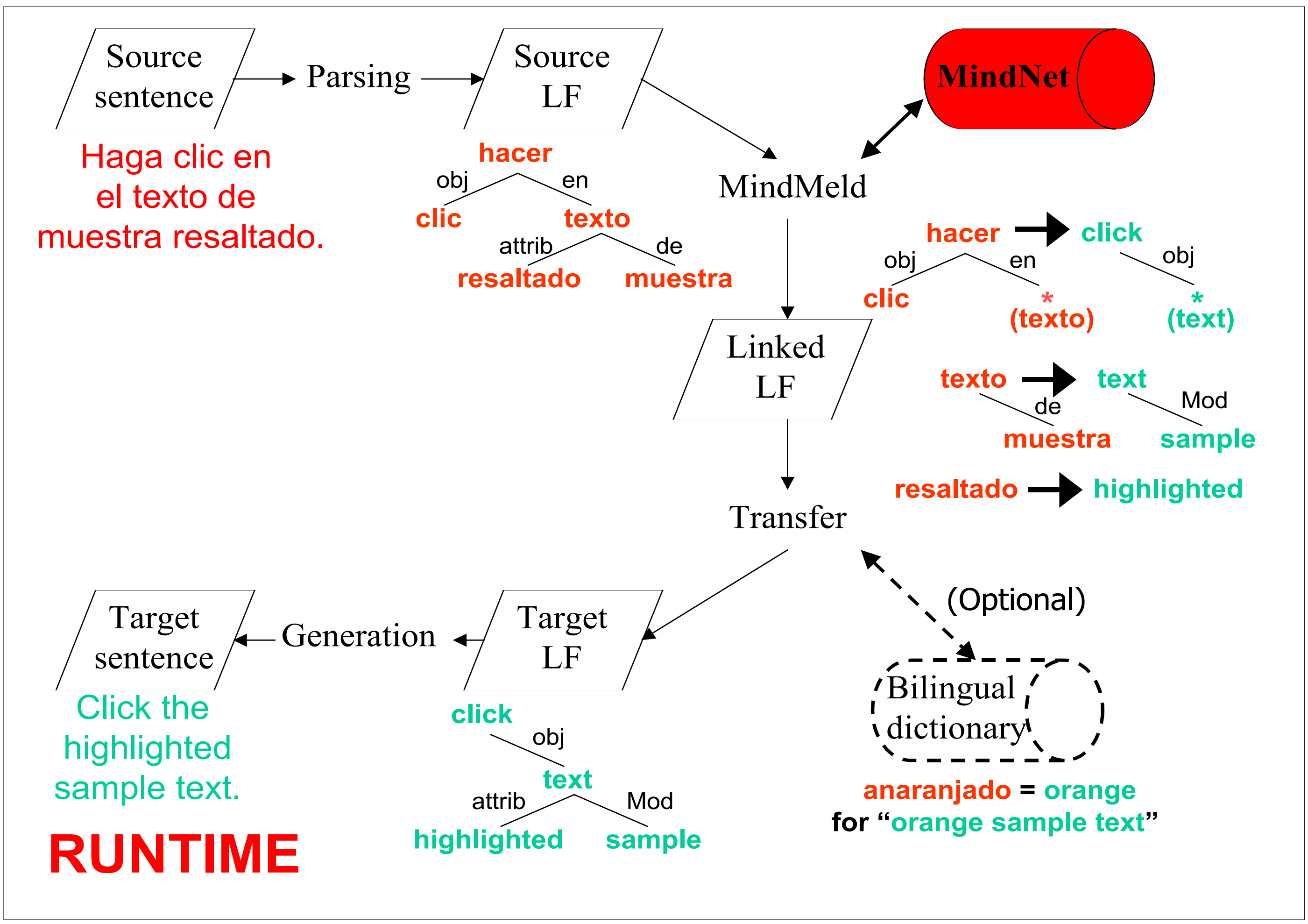


Figure 4: An example of how an english string "Click the highlighted sample text" is generated from the original Spanish input "Haga clic en el texto de muestra resaltado."

with the results obtained using the translated Spanish PSS Knowledge Base.

Future

What is next step after the NLP Group finishes the English to Japanese, German, French, and Chinese versions of NLPWin? Dolan and Richardson see Microsoft packaging their MT technology for sale to other large corporations that need to translate huge bodies of documents. Beyond that, the NLP Group hopes to eventually have their MT system included in a future release of Microsoft Office.

However, creating accurate open-ended Machine Translation for this popular productivity suite presents a major problem. The system must be capable of translating languages across many domains, with sparse data, which is the opposite of the PSS Knowledge Base translation, where the domain is narrow and plenty of sample data exists. Richardson believes that to overcome this obstacle will require a cooperative effort. He foresees a time when MT systems will link together across the Internet.

"In my personal opinion, the pathway where we start with Microsoft [internally] and go out to other companies [externally], will eventually lead to many different MT systems on the Internet tuned up for different domains. At some point, they will merge into a 'mega-translator,' where you send in a document, which is then classified for the domain it belongs to, and is passed along to a particular MT system tuned up for that domain, which then will do a reasonable job of translating it," says Richardson.

Microsoft is notorious for developing lucrative software. Will this mega-translator concept be another such source of income? According to Richardson, probably not. "It's like email—MT is so essential to communication and will become so ubiquitous that no one will make big money with it. Free lower quality MT already exists out on the Internet. Initially, self-customizing, higher quality MT systems will make money, but in the long run, high quality MT will probably also be free," says Richardson.

Don Barker is Senior Editor of PC AI Magazine, and author of twenty-four computer textbooks. Some of the material in this article was gathered with Stuart J. Johnston for a book.

URLs Related to this Article

The NLP Group site is:
http://research.microsoft.com/nlp

IBM Watson Labratory:
www.watson.ibm.com/facility_history.html

Cyc Incorporated:
www.cyc.com

Other related URLs

The Natural Language Processing FAQ:
www-2.cs.cmu.edu/Groups/AI/html/faqs/ai/nlp/nlp_faq/faq.html

The Association of Computational Linguistics:
www.aclweb.org

Natural Language Processing and Computational Linguistics
www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp.html