Role of Machine Translation in internationalizing and localizing web based product

Manoj Mulik*,

Assistant professor and Head Computer Engg, Anantrao Pawar College of Engineering and Research, Parvati, Pune

*manojmulik@gmail.com

***Abstract*— Translation of a web site is one of the most effective methods to secure the reputation in the international markets. According to statistics more than 50% of the Internet users either live outside the English-speaking countries or do not speak English at all. The scientific research in the sphere of information technologies also reveals that there is a greater chance to make a purchase if the site address of potential clients is in their mother tongue. Web site translation gives the opportunity to attract not only English-speaking customers, but also all those who speak other native languages. Translation of web site is a complex operation which requires skills, experience and professional knowledge in many areas. We propose a system which includes two main tasks -a professional and grammatically correct translation of the website content, and an adaptation of the web site navigation, graphics, optimization and other components. Generally the translation process is carried out in two steps. First the translator mechanism needs to detect possible modifications and flaws in the original text and understand the meaning that is to be conveyed. In second step the translator mechanism will unwind the syntactic structure of the original text and then formulate the corresponding message in the target language, thus giving the original text added value in terms of both wording and impact. In these ongoing efforts we intend to have a translation system which will provide more accurate translation than currently available tools such as google translate. We will also develop a web application which will be automatically internationalized or localized according to the user's locale.**

***Keywords*- Machine Translation, (MAHT) Machine aided human translation, computer aided translation, Source Language, Target Language.**

I. INTRODUCTION

The translation of natural languages by machine, first dreamt of in the seventeenth century, has become a reality in the late twentieth. Computer programs are producing translations - not perfect translations, for that is an ideal to which no human translator can aspire; nor are translations of literary texts, for the subtleties and nuances of poetry beyond computational analysis; but translations of technical manuals, scientific documents, commercial prospectuses, administrative memoranda, medical reports [1]. Machine translation is not primarily an area of abstract intellectual inquiry but the application of computer and language sciences to the development of systems answering practical needs. After an outline of basic features, the history of machine translation is traced from the pioneers and early systems of the 1950s and 1960s, the impact of the ALPAC report in the mid-1960s, the revival in the 1970s, the appearance of commercial and operational systems in the 1980s, research during the 1980s, new developments in research in the 1990s, and the growing use of systems in the past decade. This brief history can mention only the major and most significant systems and projects [13].

II. BASIC FEATURES AND TERMINOLOGIES

The term 'machine translation' (MT) refers to computerized systems responsible for the production of translations with or without human assistance. It excludes computer-based translation tools which support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, etc. The boundaries between machine-aided human translation (MAHT) [12] and human aided machine translation (HAMT) are often uncertain and the term computer-aided translation (CAT) can cover both, but the central core of MT itself is the automation of the full translation process. Although the ideal goal of MT systems may be to produce high-quality translation, in practice the output is usually revised (post-edited). It should be noted that in this respect MT does not differ from the output of most human translators which is normally revised by a second translator before dissemination [11].

However, the types of errors produced by MT systems do differ from those of human translators (Incorrect propositions, articles, prônons, verb, tenses, etc.). Post-editing is the norm, but in certain circumstances MT output may be unedited or only lightly revised, e.g. if it is intended only for specialists familiar with the text subject. Output might also serve as a rough draft for a human translator, i.e. as a 'pre-translation'.

The translation quality of MT systems may be improved either, most obviously, by developing more sophisticated methods or by imposing certain restrictions on the input. The system may be designed, for example, to deal with texts limited to the sublanguage (vocabulary and grammar) of a particular subject field (e.g. biochemistry) and/or document type (e.g. patents). Alternatively, input texts may be written in a controlled language, which restricts the range of vocabulary, and avoids homonymy and polysemy [5] and complex sentence structures. A third option is to require input texts to be marked (pre-edited) with indicators of prefixes, suffixes, word divisions, phrase and clause boundaries, or of different grammatical categories (e.g. the noun convict and its homonymous verb convict).

III. DESIGN OF A MACHINE TRANSLATION SYSTEM

Systems are designed either for two particular languages (bilingual systems) or for more than a single pair of languages (multilingual systems). Bilingual systems may be designed to operate either in only one direction (unidirectional), e.g. from Japanese into English, or in both directions (bidirectional). Multilingual systems are usually intended to be bidirectional; most bilingual systems are unidirectional [6].

In overall system design, there have been three basic types. The first (and historically oldest) type is generally referred to as the 'direct translation' approach: the MT system is designed in all details specifically for one particular pair of languages, e.g. Russian as the language of the original texts, the source language, and English as the language of the translated texts, the target language. Translation is direct from the source language (SL) text to the target language (TL) text; the basic assumption is that the vocabulary and syntax of SL texts need not be analysed any more than strictly necessary for the resolution of ambiguities, the correct identification of TL expressions and the specification of TL word order; in other words, SL analysis is oriented specifically to one particular TL. Typically, systems consist of a large bilingual dictionary and a single monolithic program for analysing and generating texts; such 'direct translation' systems are necessarily bilingual and unidirectional [7].

IV. DESIGN STRATEGIES FOR A MACHINE TRANSLATION SYSTEM

The basic design strategy is the Interlingua approach, which assumes that it is possible to convert SL texts into representations common to more than one language. From such Interlingua representations texts are generated into other languages. Translation is thus in two stages: from SL to the Interlingua (IL) and from the IL to the TL. Procedures for SL analysis are intended to be SL-specific and not oriented to any particular TL; likewise programs for TL synthesis are TL-specific and not designed for input from particular SLs. A common argument for the Interlingua approach is economy of effort in a multilingual environment. Translation from and into n languages requires n (n-1) bilingual 'direct translation' systems; but with translation via an Interlingua just 2n Interlingua programs are needed. With more than three Languages the Interlingua approach is claimed to be more economic. On the other hand, the complexity of the Interlingua itself is greatly increased. Interlinguas may be based on an artificial language, an auxiliary language such as Esperanto, a set of semantic primitives presumed common to many or all languages, or a 'Universal' language-independent vocabulary [20].

Another basic strategy is the less ambitious transfer approach. Rather than operating in two stages through a single Interlingua representation, there are three stages involving underlying (abstract) representations for both SL and TL texts. The first stage converts SL texts into abstract SL-oriented representations; the second stage converts these into equivalent TL-oriented representations; and the third generates the final TL texts. Whereas the Interlingua approach necessarily requires complete resolution of all ambiguities in the SL text so that translation into any other language is possible, in the transfer Approach only those ambiguities inherent in the language in question are tackled; problems of lexical Differences between languages are dealt with in the second stage (transfer proper). Transfer systems consist typically of three types of dictionaries (SL dictionary/ies containing detailed morphological, Grammatical and semantic information, similar TL dictionary/ies, and a bilingual dictionary relating base SL forms and base TL forms) and various grammars (for SL analysis, TL synthesis and for transformation of SL structures into TL forms)[2].

V. DIRECT APPROACH

The direct approach lacks any kinds of intermediate stages in translation processes: the processing of the source language input text leads 'directly' to the desired target language output text. In certain circumstances the approach is still valid today — traces of the direct approach are found even in indirect systems — but the first direct MT systems had a more primitive software design. A direct MT system is designed in all details specifically for one particular pair of languages in one direction, e.g. Russian as the language of the original texts, the source language, and English as the language of the translated texts, the target language. Source texts are analysed no more than necessary for generating texts in the other language.

The direct approach is summarized in the figure below

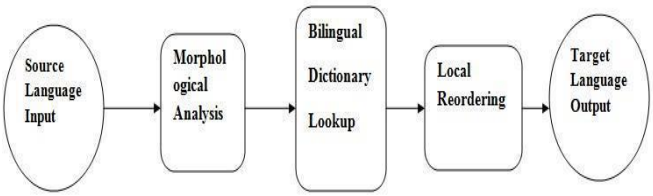

Fig. 1 The direct approach

First generation direct MT systems began with what we might call a morphological analysis phase. In this phase the system identified word endings and reduced inflected forms to their uninflected basic (canonical) forms. Then it input the results into a large bilingual dictionary look-up program. There would be no analysis of syntactic structure or of semantic relationships! In other words, when the system would find the canonical form of a word, it would look it up in the bilingual dictionary to find an equivalent in the target language. There would follow some local reordering rules to give more acceptable target language output, perhaps moving some adjectives or verb particles, and then the target language text would be produced[18].

VI. INTERLINGUA APPROACH

The first is the Interlingua method where the source text is analysed in a representation from which the target text is directly generated. The intermediate representation includes

all information necessary for the generation of the target text without 'looking back' to the original text. This is an abstract representation of the target text as well as a representation of the source text. It is neutral between two or more languages. In the past, the intention or hope was to develop a representation which was truly 'universal' and could thus be intermediary between any natural languages. At present, interlingual systems are less ambitious. The Interlingua approach is clearly most attractive for multilingual systems. Each analysis module can be independent, both of all other analysis modules and of all generation modules [16].

Target languages have no effect on any processes of analysis; the aim of analysis is the derivation of an 'interlingual' representation. The advantage is that to add a new language to the system one needs to create just two new modules: an analysis grammar and a generation grammar. There are major disadvantages to the interlingual approach. The main is the difficulty of creating an Interlingua, even for closely related languages (e.g. the Romance languages: French, Italian, Spanish, and Portuguese). A truly 'universal' and language-independent Interlingua hasn't been created so far [15].

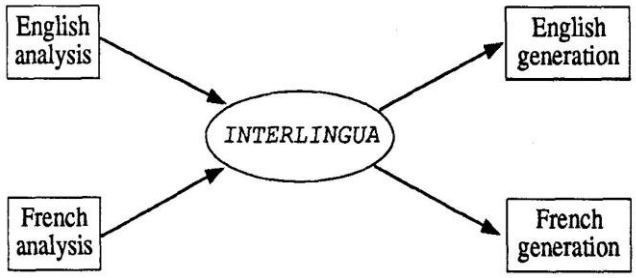

Fig. 2 Interlingua representation

The above diagram represents an Interlingua for the translation of text from English to French.

VII. TRANSFER SYSTEMS

The second variant of the indirect approach is called the transfer method. Although there is some kind of 'transfer' in any translation system, the term transfer method applies to those which have bilingual modules between intermediate representations of each of the two languages. These representations are language-dependent: the result of analysis is an abstract representation of the source text (this could be something like a phrase-structure tree). In turn, the input to generation is an abstract representation of the target text (again, possibly a tree). The function of the bilingual transfer modules is to convert source language (intermediate) representations into target language (intermediate) representations. Since these representations link separate modules (analysis, transfer, generation), they are also frequently referred to as interface representations [9].

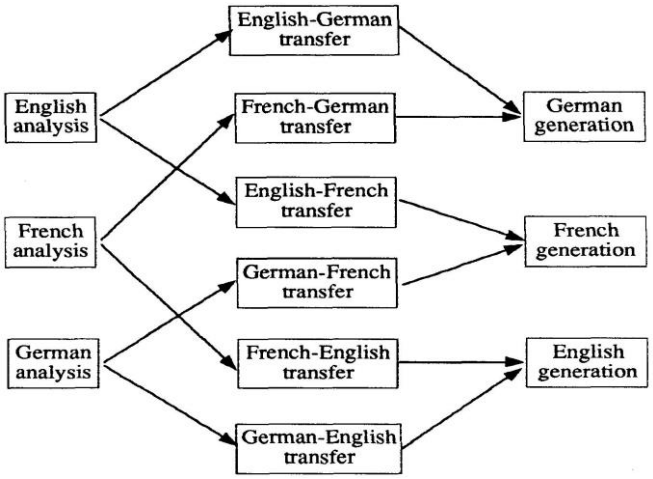

Fig. 3 Transfer System

The above figure shows conversion of source language modules (intermediate) to source language modules (intermediate) which is used in transfer systems.

VIII. MACHINE TRANSLATION

Machine Translation (MT) is the task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language. While machine translation is one of the oldest subfields of artificial intelligence research, the recent shift towards large-scale empirical techniques has led to very significant improvements in translation quality. The term machine translation (MT) is used in the sense of translation of one language to another. The ideal aim of machine translation systems is to produce the best possible translation without human assistance. Basically every machine translation system requires programs for translation and automated dictionaries and grammars to support translation. The translation quality of the machine translation systems can be improved by pre-editing the input. Pre-editing means adjusting the input by marking prefixes, suffixes, clause boundaries, etc. Translation quality can also be improved by controlling the vocabulary. The output of the machine translation should be post-edited to make it perfect. Post-editing is required especially for health related information. Machine translation is one of the research areas under "computational linguistics". Various methodologies have been devised to automate the translation process. However, the objective has been " to restore the meaning of original text in the translated verse". In general, the process of translation has two levels, Metaphase and Paraphrase [19].

Metaphrase means "word-to-word" translation. It relates to "formal equivalence", i.e., the translated version will have "literal" translation for each word in the text. However, the translated text may not necessarily convey the meaning of the original text. That means sometimes the semantics may differ from the original text.

Paraphrase relates to "dynamic equivalence", i.e., the translated text would contain the gist of the original text

but may not necessarily contain the word-to-word translation.

A. *Types of Machine Translation System*

Following are the types of the Machine Translation System 1) Bilingual Systems 2) Multilingual Systems

1) Bilingual System

Machine translation systems that produce translations between only two particular languages are called bilingual systems. Bilingual Systems can be unidirectional and bidirectional..Bilingual Systems can be reversible or non-reversible. Unidirectional means the translation from one language to another language is possible only in one direction i.e. (Language1 → Language2)[18].

Bidirectional means the translation from one language to another language and vice versa is possible in both directions i.e. (Language1 ↔ Language2).

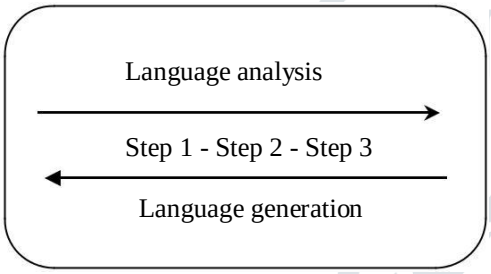

Fig. 4 Language analysis and generation

In reversible system (see the above Fig 4) the process of language generation is the opposite of language analysis. For example, the English analysis module in an (English → German) system will mirror the English → German generation module in a German → English system.However; it's too difficult to design a truly reversible bilingual system, so nearly all bilingual systems are in effect two uni-directional systems running on the same computer. Such a bilingual system is best represented by the following means. (Language1 → Language2)+(Language1 ← Language2) instead of Language1 ↔ Language2.

Methods of analysis and generation for either of the languages are designed independently. A bilingual system is therefore, typically, one designed to translate from one language into one other in a single direction.

2) Multilingual System

Machine translation systems that produce translations for any given pair of languages are called multilingual systems. Multilingual systems are preferred to be bi-directional and bi-lingual as they have ability to translate from any given language to any other given language and vice versa[8].

B. *Basic approaches for machine translation*

Following are the Basic approaches used for the machine translation 1) Direct Machine Translation Approach.2) Interlingua Approach.3) Transfer Approach.

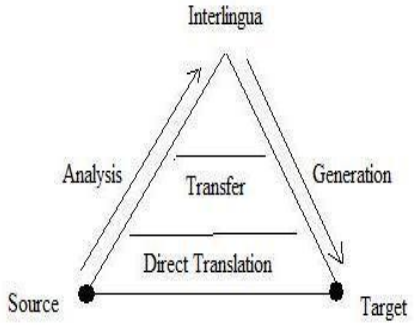

Fig. 5 Machine Translation Pyramid

1. Direct Machine Translation Approach

Direct translation approach is the oldest and less popular approach. Machine translation systems that use this approach are capable of translating a language, called source language (SL) directly to another language, called target language (TL). The analysis of SL texts is oriented to only one TL. Direct translation systems are basically bilingual and uni-directional. Direct translation approach needs only a little syntactic and semantic analysis. SL analysis is oriented specifically to the production of representations appropriate for one particular TL [4].

2. Interlingua Approach

Following diagram demonstrates the how machine translation is done using Machine Translation.

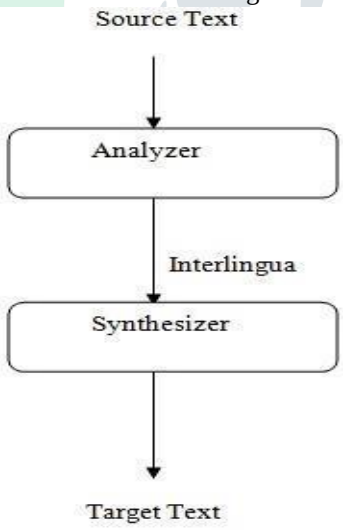

Fig 6 Interlingua Machine translation

This is considered to belong to third generation of machine translation. It is an inherent part of a branch called Interlinguistics.Interlingua aims to create linguistic homogeneity across the globe. Interlingua is a combination of

two Latin words Inter and Lingua which means between/intermediary a n dlanguage respectively. In Interlingua, source language is transformed into an auxiliary/intermediary language (representation) which is independent of any of the languages involved in the translation. The translated verse for the target language is then derived through this auxiliary representation. Hence, only two modules i.e., analysis and synthesis are required in this type of system. Also, because of its independency on the language pair for translation, this system has much relevance in multilingual machine translation. This emphasizes on single representation for different languages. The parameterization model proposed by Ali is one of the enhancements over inter-lingua model with only one analysis component (multi-lingual parser) and one synthesis which work multilineally.The UNITRAN system is one implementation of this model [13].

Interlingua approach intends to translate SL texts to that of more than one language. Translation is from SL to an intermediate form called Interlingua (IL) and then from IL to TL. Interlingua may be artificial one or auxiliary language like Esperanto with universal vocabulary. Interlingua approach requires complete resolution of all ambiguities in the SL text. Interlingual Machine Translation is a methodology that employs Interlingua for translation. Ideally the interlingual representation of the text should be sufficient to generate sentences in any language. Languages can have different parts of speech. In some cases two or more words in one language have an equivalent single word in another language. Interlingua approach addresses these structural differences between languages. The disadvantage is that the design of Interlingua is too complex. This is due to the fact that there is no clear methodology developed so far to build a perfect interlingual representation. An interlingual lexicon is necessary to store information about the nature and behavior of each word in the language. The information includes events and actions. A typical interlingual MT system has analyzer and synthesizer for each language. The analyzer produces interlingual representation of the meaning of the given text. The synthesizer produces one or more sentences with the meaning given by the analyzer. KANT[3] is an interlingual machine translation system operational in a commercial setting .

3. Transfer Approach

Unlike Interlingua approach, transfer approach has three stages involved. In the first stage, SL texts are converted into abstract SL-oriented representations. In the second stage, SL-oriented representations are converted into equivalent TL-oriented representations. Final texts are generated in the third stage. In transfer approach complete resolution of ambiguities of SL text is not required, but only the ambiguities inherent in the language itself are tackled. Three types of dictionaries are required: SL dictionaries, TL dictionaries and a bilingual transfer dictionary. Transfer systems have separate grammars for SL analysis, TL analysis and for the transformation of SL structures into equivalent TL forms. Transfer model belongs to the second generation of machine translation (mid 60s to 1980s). In this, source language is transformed into an abstract; less language-specific representation. An equivalent representation (with same level of abstraction) is then generated for the target language using bilingual dictionaries and grammar rules. These systems have three major components [10].

i.Analysis

Analysis of the source text is done based on linguistic information such as morphology, part-of-speech, syntax, semantics, etc. Heuristics as well as algorithms are applied to parse the source language and derive the syntactic structure (for language pair of the same f a m i l y , f o r e x a m p l e T a m i l a n d Telugu are siblings of same family i.e. Dravidian Languages etc.) of the text to be translated. Or the semantic structure (for language pair of different families, Hindi from Devnagari Family and Telugu from Dravidian Family).

ii.Transfer

The syntactic/semantic structure of source language is then transferred into the syntactic/semantic structure of the target language [14].

iii.Synthsys

This module replaces the constituents in the source language to the target language equivalents. This approach, however, has dependency on the language pair involved. Thus, two independent monolingual dictionaries were suggested in Eurotra project. Also, there are different representations for different languages. PaTrans (Translation for Patents) is based on transfer based approach and is one of the outcomes of Eurotra Research. Mantra is also a translation m o d e l f o r I n d i a n Languages based on transfer approach. It is Government of India funded project and the parser used for language processing is known as Vyakarta [3].

IX. PROPOSED WORK

Authors are working in the direction to make a proper translation system. This system will give the proper translation by studying the semantic and syntactic structure of the language and by writing more accurate grammar rules for the translation. Also our aim is to apply automatic localization/internationalization to the webproduct depending upon user's locale.

X. CONCLUSIONS

By following the Interlingua approach for language translation one can have an intermediate representation called 'Interlingua'. As this Interlingua is neither related to the source language nor to the target language and it is an abstract representation of the source language. It is easier to generate many target languages from one source language with the help of the Interlingua. We have reported our work of creating a new and improved translation system.

REFERENCES

[1] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in Proc. Meeting Assoc. Comput. Linguist., 1991, pp. 177–184.

[2] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer, "The mathematics S. Vogel, H. Ney, and C. Tillmann, "HMM based word alignment in Statistical translation," in *Proc. COLING*, 1996, pp. 836–841.

[3] K. Toutanova, H. T. Ilhan, and C. Manning, "Extentions to HMMbased statistical word alignment models," in *Proc. EMNLP*, 2002, pp. 87–94.

[4] M. Epstein, K. Papineni, S. Roukos, T.Ward, and S. D. Pietra, "Statistical natural language understanding using hidden clumpings," in Proc.ICASSP, Atlanta, GA, May 1996, vol. 1, pp. 176–179.

[5] M. Epstein, "Statistical source channel models for natural language understanding," Ph.D. dissertation, New York Univ., New York, Sep. 1996.

[6] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition,"

[7] S. D. Pietra, M. Epstein, S. Roukos, and T. Ward, "Fertility models for statistical natural language understanding," in Proc. 8th Conf. Eur. Chapt. Assoc. Comput. Linguist., Morristown, NJ, 1997, pp. 168–173, Association for Computational Linguistics.

[8] Meeting on Association for Computational Linguistics, Morristown, NJ, USA, 1991, pp. 264–270, Association for Computational Linguistics.

[9] A. L. Berger, S. D. Pietra, and V. J. D. Pietra, "A maximum entropy approach to natural language processing," Comput. Linguist., vol. 22, no. 1, pp. 39–71, 1996.

[10] F. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in Proc. Joint Conf. Empirical Methods Natural Lang. Process. Very Large Corpora, College Park, MD, 1999, pp. 20–28.

[11] Using Parallel Texts: Data Driven Machine Translation and Beyond, R. Mihalcea and T. Pedersen, Eds., Edmonton, AB, Canada, May 31, 2003, pp. 73–80, Association for Computational Linguistics.

[12] F. Och, "Statistical machine translation: From single word models to alignment templates," Ph.D. dissertation, RWTH Aachen, Aachen, Germany, 2002.

[13] "The NIST machine translation evaluations." NIST, 2004 [Online]. Available: http://www.nist.gov/speech/tests/mt/
"LDC Chinese Segmenter." LDC, 2002 [Online]. Available: http://www.ldc.upenn.edu/Projects/Chinese

[14] Y. Deng, S. Kumar, and W. Byrne, "Segmentation and alignment of parallel text for statistical machine translation," J. Natural Lang. Eng., vol. 13, no. 3, pp. 235–260, 2006.

[15] of machine translation: Parameter estimation," *Comput. Linguist.*, vol. I. D. Melamed, "Models of translational equivalence among words," *Comput. Linguist.*, vol. 26, no. 2, pp. 221–249, 2000.

[16] D. Marcu and W. Wong, "A phrase-based, joint probability model for Statistical machine translation," *Proc. EMNLP*, pp. 133–139, 2002.

[17] Y. Zhang and S.Vogel, "An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora," in *Proc. 10th Conf. Eur.* Assoc. Mach. Translation, 2005, CD-ROM.

[18] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, 2003.

[19] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. HLT-NAACL*, 2003, pp. 127–133.

[20] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "Wordsense disambiguation using statistical methods," in Proc. 29th Annu.