Machine Translation by Semantic Features

by
Uzzi Ornan and Israel Gutter
Computer Science Department,Technion – Israel Institute of Technology

Abstract

This article presents a method for analyzing a sentence and translating it by means of identifying its various semantic components as complements of the main verb. Order of words of the source language, which may be quite free, becomes unimportant in this analysis and translation process. In source languages with a deficient writing system, such as Hebrew or Arabic, this process eliminates misreading of the strings of letters (words). Our project is operating on Hebrew as its source language. This language has rich morphology with multiplicity of meaning, and the order of the words in it is fairly free.

1. Why translating from Hebrew is so confusing?

Mechanical translation from Hebrew into other languages is considered a very difficult process. The reason is the deficient writing system of Hebrew. It does not show most of the vowels, never uses two identical letters to signify a gemination, does not separate several particles from the following word, and sometimes one and the same character may indicate either a vowel or a consonant. Another feature of Hebrew is its rich inflection of verbs, nouns and prepositions. Inflection changes person, number and gender of the subject of the verb, as well as the object of a verb – which may be incorporated as a suffix of a verb. Inflection changes the pronoun when it is attached to a preposition. It may also change the pronoun suffix of a noun indicating the owner. Inflection in Hebrew may generate twenty to forty different morphological structures for each word (according to other approach it may reach several hundreds). This richness adds, even more, to the high number of interpretations for a Hebrew string of letters.

The result is that each string of characters may have more than one reading. The average number of possible readings for each string of letters is almost three. This graphemic situation makes computer processing of a Hebrew phrase or clause rather complicated. For example, if we have the following string consists of four letters, LBNH , there are more than seven possible readings: lbana (moon), libne (styrax tree), lbena (brick), labbne (yogurt),), labbnah (make her clear), libbnah (made her clear), l-bnah (to her son). Beside these readings, some of them may have more than one meaning, such as lbana, that may have another meaning: adjective "white" (feminine) besides "moon".

Richness of grammatical forms in a language makes fix order of words unnecessary,

since very easily listeners can relate verbs to their subjects, as well as other connections between words. Indeed order of words is rather free in Hebrew[1]. Thus, if we have a sentence of five words, each of them having three readings, assuming that no fixed order of words is necessary, the calculated number of possible sentences may reach 243 ($= 3^{5)}$. Most of these, however, are not sentences at all. Human beings choose the proper reading – in most cases - of each string by looking at the syntax, semantics and pragmatics. But what about Hebrew processed by a computer? We must assume our program should include syntax, semantics and world knowledge.

2. Existing Processing of Hebrew

At present, most attempts of processing Hebrew, which must be the first step towards translating Hebrew texts into other languages, does not include syntax or semantics. The main efforts have been made in morphology. At IBM Israel Scientific Center a project has been compiled in which each Hebrew string of characters gets all possible morphological analyses (Bentor et al. 1992). Similar projects have been done by others[2], but deciphering Hebrew sentences heavily rely on statistical and heuristic considerations[3], on top of "short context" syntax, i.e. two or three neighboring words are checked together to see whether they do not contradict each other[4,5]. Another restricting approach to the problem is that usually, English order of words is assumed as prevailing in Hebrew texts also. As far as we know, there is no program for translating full sentences from Hebrew into other languages.

3. Translating According to Case Theory

In order to overcome the difficulties of processing Hebrew we adopted the case theory approach, in which syntactic valency of verbs is also used. This method is mentioned in most introductions to Machine Translation[6], but is usually criticized as not sufficient for full processing. Still, many researchers include some variety of case theory in their projects (For example, Schank-Rieger(1974:381), Wilks(1975:62f)). Somers, in his wide, comprehensive monograph (Somers 1987), describes several projects for MT, most of them using Case Grammar together with other approaches for parsing and translating. Following them we have chosen Case approach as the main part of our project, since we consider it solving the problem of free order of words, which is so important in Hebrew.

For example, all the following six sentences, literally translated from all combinations of "sus ʔakal ʿeśb" (= horse ate grass), are possible:
A horse ate grass - A horse grass ate - Ate a horse grass - Ate grass a horse - Grass a horse ate - Grass ate a horse.
All these sentences are more particularly acceptable and right-sounding if 'horse' (the performer of the act) is preceded by a definite article ("ha-sus").

This article describes our project: we analyze a sentence in the source language (Hebrew) by treating its main verb as a function (or mathematical predicate). Nominal phrases (NP) and prepositional phrases (PP) are taken as its arguments. We have built a detailed lexicon that is divided into two main parts. The first lexicon is constructed of verbs as entries. Here each entry includes the expected thematic roles of NP's (PP's included) of the sentence containing this verb, and for each of them a list of expected semantic features is provided[7]. The second lexicon is that of the noun. Here semantic features of the signified item are given. Each meaning of a verb or a noun occupies a separate entry. Other shorter lexicons are of adjectives and adverbs. Prepositions are incorporated in the verb lexicon. At present the lexicons contain about 30,000 entries.

3. Examples from our lexicons.

For this article we have introduced some changes in the original disposition. Firstly we replaced our Hebrew names for thematic roles and semantic features with English ones. Thematic roles are given in capital letters. Curly brackets contain semantic features. Slash means exclusive "or" (either what is before the slash or what is after it). Parentheses indicate optional components. English translation of the Hebrew entry is given in the same line preceded by an asterisk. Hebrew words (mainly prepositions, or parts of idioms) are enclosed in inverted commas. Latin characters for Hebrew alphabet are according to ISO/FDIS/259-3.

A. From the verb lexicon:

hebin *understand
NP1 EXPERIENCER {human, role, org.}
NP2 "še-" %SENTENCE
/
NP2 "et" AIMED-AT {abstract, info}

šimmeš1 *serve
NP1 INFLUENCER {-}
NP2 THEME {human, org.}
(NP3) {"l-"/"k-"} GOAL {action}

šimmeš2 *use
NP1 THEME {human, instrument, site, construction}
NP2 "btor" FUNCTION { human, instrument, site, construction}

ʕarak1 *set
NP1 AGENT {human}
NP2 "et" THEME {"šulxan"} (=table. An idiom)
ʕ&arak2 *hold
NP1 AGENT {human, org.}
NP2 "et" THEME {act, happening}

ʕarak3 *edit
NP1 AGENT {human, org.}
NP2 "et" THEME {printed_matter}

zaraḥ *rise
NP1 THEME {source_of_light, source_of_heat, strong}

rakab *ride
NP1 AGENT {human}
(NP2) "ʕal" THEME {four_legged_animal, vehicle}
(NP3) "l-" TO-LOC {site, place, happening, human}
(NP4) "mi-" FROM- LOC {site, place, happening, human}

B. **From the noun lexicon**:

šammaš *caretaker {human, function}
šemš *sun {source_of_heat, source_of_light, strong, periodical}
šulḥan *table {furniture, utensil}
yarḥon *monthly {printed_matter, periodical}
boqr *morning {time}
boqer *cowboy {human, male,function}
ner *candle {source_of_heat, source_of_light, weak}
ḥtunna *wedding {happening}
ḥamma *sun {source_of_heat, source_of_light, strong, periodical}
ḥema *anger (feeling, mood}
yeld *boy {human, young, male}

C. **From the adjective lexicon**
bahir *clear
yape *nice

ḥazaq *strong
ḥamm *hot
Adjectives change according to gender and number.

4. The program

The program reads each Hebrew string, and produces analyses of all possible readings for that string. For example, giving the sentence ha-boqr zarḥa šemš ḥamma (given in Hebrew script – הבוקר זרחה שמש חמה "hot sun rose this morning") will render the following:

hbwqr	N boqr	,a,-,- ,-,3,+,#,s	-,-,-,-,- ha-
	N boqer	,a,-,- ,-,3,+,#,s	-,-,-,-,- ha-
zrḥh	N zarḥa	,a,-,- ,-,3,#,+,s	-,-,-,-,-
	N zerḥ	,i,-,- ,-,3,+,#,s	3,#,+,s,h
	V zaraḥ	,-,-,h ,p,3,#,+,s	-,-,-,-,-
šmš	N šammaš	,a,-,- ,-,3,+,#,s	-,-,-,-,-
	N šammaš	,c,-,- ,-,3,+,#,s	-,-,-,-,-
	N šemš	,a,-,- ,-,3,#,+,s	-,-,-,-,-
	N šemš	,c,-,- ,-,3,#,+,s	-,-,-,-,-
	A šammaš	,a,-,- ,-,-,+,#,s	-,-,-,-,-
	V šimmeš	,-,-,- ,i,2,+,#,s	-,-,-,-,-
	V maš	,-,-,- ,p,3,+,#,s	-,-,-,-,- $e-
	V maš	,-,-,- ,r,-,+,#,s	-,-,-,-,- $e-
ḥmh	N ḥema	,a,-,- ,-,3,#,+,s	-,-,-,-,-
	A ḥamma	,a,-,- ,-,-,#,+,s	-,-,-,-,-
	N ḥamma	,a,-,- ,-,-,#,+,s	-,-,-,-,-

The given Hebrew string is in the first column. Second and third columns are the category and the lexical entry. Other columns give details of morphological features. Last column specifies attached particles. The program is looking first for a verb. When a verb is identified, the program checks the lexicons in order to see whether the NP's accord with the expected thematic role of the proposed verb, and whether they contain the appropriate semantic features.

This procedure is repeated for each possible reading of a string as a verb in order to discover all possible interpretations of the sentence. Then the program turns to the target language. Since each meaning of each verb and each noun is listed as a separate entry in the lexicon, an adequate translation for each word into the target language could be provided in the lexicon. Moreover, several translations into various languages may be included in the same source entry. Very easily, with minimal effort, we thus achieve a multi-lingual translator originating from the same lexicon. In order to generate a proper phrase or clause in the target language, we include a generator for each target language.

Our target language at present is English, but as one can see in the following examples, some work has been done on Spanish too.

This practice may have a theoretical justification: let us look at the entries of the lexicons as "concepts" rather than "words". What in fact exist there – Hebrew words – are just a device to signify these universal concepts, and easily can be replaces by words in another language. Admittedly, this approach cover the lexicons only, and does not include necessary details in the structure of languages, such as agreement, order, prepositions etc..

5. **Examples** (processed by Dan Yaacobi and Meirav Greenberg, Technion lab of 1998-9)

(1) Hebrew Sentence: bhtwnh ʕrk hyld šwlhn
Interpretation: bḥtwnh(Na '-b-'ḥtunna f3s wedding)[happening] hyld(Na '-ha-'yeld m3s boy)[human,young,male] ʕrk(v '-'ʕarak1 m3sp-a set) šwlḥn(Na '-'šulḥan m3s table)[utensil,furniture]
English: **The boy set a table at a wedding**.

(2) Hebrew Sentence: hyld ʕrk šwlḥn
Interpretation: hyld(Na '-ha-'yeld m3s chico)[human,young,male] ʕrk(v '-'ʕarak1 m3sp-a puso) šwlxn(Na '-'šulḥan m3s mesa)[utensil,furniture]
Spanish: **El chico puso un mesa .**

(3) Hebrew Sentence: hw^ ʕrk yrḥwn
Interpretation: hw^(p '-'hu^ m3s el)[human] ʕrk(v '-'ʕarak3 m3sp-a edito') yrḥwn(Na '-'yarḥon m3s revista mensual)[printed_matter, periodical]
Spanish: **El edito' un revista mensual .**

(4) Hebrew Sentence: hyld hbyn hḥtwnh nʕrkh.
Interpretation: hyld(Na '-ha-'yeld m3s boy)[human,young,male] hbyn(v '-'hebin m3sp-a understand) šḥḥtwnh(Na '-še-ha-'ḥtunna f3s wedding)[happening] nʕrkh(v '-'ne&rak f3sp-p hold) (a compound sntence)
English :**The boy understood that the wedding was held**.

6. **Conclusions**
One can recognize the influence of case theory, especially that of C.C. Fillmore, on our project, even though we had to manipulate it in some aspects, such as create some new thematic roles as well as enlarging the list of semantic features. Also we have made more complicated dependencies in certain cases, such as a certain role is justified only if

another role includes certain semantic features. This part is independent of any syntactic properties.

Another note is that the act of translating is just a minor part in our system. The main part is deciding what is the proper reading for each string of the Hebrew source. The first step of this procedure, as seen in (4) above, is a phonemic conversion of each reading of the Hebrew string into Latin characters, then we rule out readings that do not agree with some syntactic rules. The number of actual readings is thus diminished, and conveniently we can activate the semantic program as described above.

As mentioned above, each Hebrew string may be interpreted in an average of three ways. The main point in our engine is that it deals with a sentence rather than with a word. The result is that we achieve a precision of 80-85%, and 100% of integrity, while the regular search engines for Hebrew, which do not cope with syntax or semantics, has a precision of 20-40% only. It may be of some interest to know that the major part of the program is already in use by a commercial company that uses it mainly as a search engine. It gives its services to many institutions, such as the Knesset (Israeli parliament), schools, Government bureaus, lawyer offices etc.

This integrated method, while it is a must for Hebrew and some other languages which use a deficient writing system, such as Arabic or Persian, may also be beneficial for languages which use a full alphabet, such as languages written in Latin, Greek or Cyrillic characters. As a matter of fact, our rich lexicons, seemingly based on Hebrew only, are indeed based on universal concepts, which are expressed quite similarly in all languages.

Notes

1. According to Gűngődii & Oflazer(1995), morphological ambiguities in Turkish cause similar freedom of word order.
2. Mainly as projects in Departments of Computer Science at the Technion and other Israeli universities.
3. In Bar-Ilan Responsa project (headed by Y. Choueka) millions of running words have been gathered in frequency lists. Later these lists have been used to build a heuristic automatic program to replace non vocalized Hebrew words by vocalized ones. This program is distributed now by the Center for Educational Technology, and is claimed to have 85-90% in precision. This heuristic approach serves also as a research engine, which seemingly do not reveal these 10-15% cases.
4. Choueka (1990) uses "Short-context tools" as an interactive device.
5. Hertz and Rimon show a promising algorithm for diminishing ambiguity. Some projects for M. Sc. degree on short context have been conducted in the recent

years in Israeli universities.

6. For example, see Hutchins 1986, Hutchins and sommers 1992, Arnold et al. 1994.
7. One should recall the "semantic markers" of Katz-Fodor(1964) which played a similar role in their semantic theory.

References

* Arnolds D., L.Balkan, R. Lee Humphreys, S. Meijer and L. Sadler (1994), *Machine Translation, an Introductory Guide*, Blackwell, Oxford UK.
* Bentor E., A. Angel, D. Ben-ari Segev and A. Lavie (1992), "Computerized analysis of Hebrew Words", in *Hebrew Computational Linguistics*, ed. by U.Ornan, G. Arieli and E. Doron, Ministry of Science and Technology, Jerusalem, Israel, p. 36-38..
* Choueka Y. (1990), "Responsa: an operational full-text retrieval system", in *Computers in Literary and Linguistic Research* ed. by J.Hamesse and A. Zampoli, Champion-Slatkine Paris – Geneve, p.87-88.
* Fillmore, C.C.(1968), "The Case for Case", in *Universals in Linguistic Theory*, ed. by E. Bach and R. Harms, Holt, Rinehart and Winston, New york, p. 1-90.
* Gűngődii, Zelal and Kemal Oflazer, (1995) "Parsing Turkish using LFG Formalism", Machine Translation, vol. 10, p. 293-319.
* Hertz Y. and M. Rimon (1992) , "Diminishing Ambiguity by short-context automaton", in *Hebrew Computational Linguistics* ed. by U.Ornan, G. Arieli, E. Doron, Ministry of Science and Technology, Jerusalem, Israel, p. 74-87.
* Hutchins, W.J.(1986), *Machine Translation, Past, Present, Future*, Ellis Horwood/Wiley, Chichester/New York.
* Hutchins W. J. and H.L. Somers (1992), *An Introduction to Machine Translation*, Academic Press, London.
* Katz, Jerrold J. and Jerry A. Fodor (1964), The Structure of a Semantic Theoty" in *The Structure of Language* (ed. by Fodor-Katz), p. 479-518, Prentice-Hall.
* Nirenburg S. and I. Nirenburg (1988), "A Framework for Lexical Selection in Natural Language Generation", in *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest.
* Schank, Roger C. and Charles J. Rieger III (1974), "Inference and the Computer Understanding of NL", *Artificial Intelligence* 5, p. 373-412.
* Somers, H.L. (1987), *Valency and Case in Computational Linguistics*, Edinburg University Press.
* Wilks, Yorick (1975), "A Preferential, Pattern-Seeing, Semantic for NL Inference", in *Artificial Intelligence* 6, p. 53-74.