Neural Network Based Lexicon Representation to support MT & ML

Asaad Al Hijaj
Computer Science Dept.,College of Science, Basrah University

أسعد عبد الكريم الحجاج

قسم علوم الحاسبات، كلية العلوم، جامعة البصرة.

الخلاصة

في هذا البحث تم اقتراح استخدام طريقة لتمثيل المعجم الدلالي من خلال بنية الشبكات العصبية كأساس لبناء نظام ترجمة آلية. النظام المقترح يعتمد أسلوب الارتباط ثنائي اتجاه في تمثيل المعجم الدلالي مما يسهل استخراج معاني المفردات للغتين مختلفتين. يهدف البحث كـذلك، لاسـتخدام طريقـة التمثيـل المقترحة كواجهة لدعم التعليم الآلي ML تسهل للبشر التعامل مع الكومبيوتر. مثال ذلك، إدخـال نـص قصة قصيرة بلغة ما تصف مشهدا وترجمتها إلى رسوم متحركة بالاستفادة من تقنيات الوسائط المتعـددة Multimedia. تم اعتماد تراكيب مناسبة لتمثيل قاعدة بيانات تشتمل على ملف للمفـردات وارتباطاتـه بملف الأصناف، ملف المعاني والجذور، ملف السوابق واللواحق، وأخيرا ملف لصيغ قواعـد التصـريف النحوي. هذا الأسلوب أتاح قدرة مناسبة في تمثيل البيانات وتشفيرها وكذلك في تقليص احتياجات الخزن.

Abstract

This paper suggests a method for neural network-based Lexicon representation to built up a machine translation system. The method adopts a Bi-Associative memory (BAM) for Lexicon representation that provides an easy access to the meaning of words in two different languages.
Appropriate techniques are used for database representation including wording file and its relation to classification file, meanings and roots file, suffixes and prefixes file and finally the morphological one.
Such method provides an appropriate means for data representation and coding in away that makes the process of analyzing and understanding the given texts accessible which ultimately might be utilized for machine learning. This would be done by finding appropriate analysis of the linguistic potentials of each word which finally leads to an accurate architecture of bi-direction associative memory. The outcome would be a minimized storage capacity needs gained through a retrieving process that is both handy and easy.A more optimistic speculation is to transform a written text into a visual animation by using multimedia techniques

Introduction and Overview

A long –standing goal for the field of artificial intelligence is to enable computer understanding of human languages. Much progress has been made in reaching this goal, but much also remains to be done. Before artificial intelligence systems can meet this goals, they first need the ability to parse sentences, or transform them into a representation that is more easily manipulated by computers. Several knowledge sources are required for parsing, such as a grammar, lexicon, and parsing mechanism.[1].

Natural language processing :

Natural languages processing (NLP) researchers have traditionally attempted to build these knowledge sources by hand, often resulting in brittle, inefficient systems that take many hundreds of hours to build.

The task of NLP is that of accepting inputs in a human natural language, and to transform the inputs into some sort of formal statements that are to be "meaningful" for a computer. The computer will be, therefore, able to react correctly to the given input; sometimes, the reaction will take the form of a NL "answer," i.e., the computer will use the formal representation corresponding to the analysis of the input to generate, in turn, statements in natural language.[2]

In natural languages, artificial intelligence and expert systems have been used for human knowledge representation, and then fabricating their intricate relationship to produce systems capable of understanding and drawing inference.[3] Due to the inadequacy in knowledge representation including symbolic one, studies have been directed to find more efficient techniques such as neural networks characterized by non-conditional learning.

Informational systems capable of dealing with natural languages is necessary for computational linguistic which provided a room for rereading the employees efforts in artificial intelligence.

In 1957, Noum Chomsky drew logical and mathematical outline to represent formal languages and their grammatical rules. This was a starting point for modern computational linguistic. This all lead Natural Language Processing. [4]

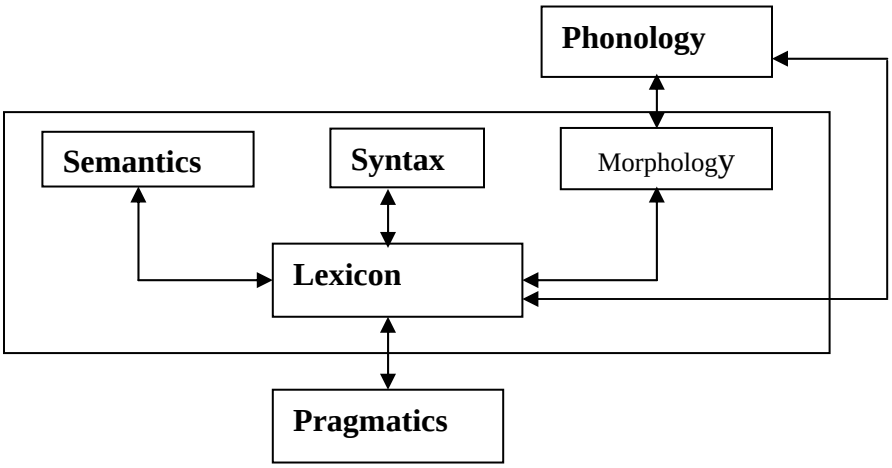

Fig (1) General Diagram for Natural Language Processing

The standard paradigm for NLP :

A comprehensive system of natural language analysis, such as an analysis module of a knowledge-based machine translation system, must include the following basic components:

o Morphological analysis: the decomposition of words into their uninflected root forms, performed at the word level. There are many morphological phenomena: almost all languages has inflectional morphology; the majority has some form of derivational morphology. A number of general models of morphological processing have been investigated. At the theoretical level, the most popular approach to morphology is the so-called two-level approach. In practical systems many other, less general and more language and task-specific approaches have been used.

o Syntactic analysis: the extraction of all well-formed syntactic structures and dependencies for a source text, performed at the sentence level. In the MT environment, a grammar must be written for each source language, in one of the many current grammar formalisms, such as, for instance, Lexical Functional Grammar, Generalized Phrase Structure Grammar, Head-driven Phrase Structure Grammar, Definite Clause Grammar, Tree-adjoining Grammar or Government-and-Binding-related Grammars. The use of a "canonical" formalism facilitates the use of a single grammar interpreter applicable to any language whose grammar is defined in the selected formalism.

o Semantic analysis: the creation of the knowledge structures in a text-meaning representation language (interlingua in MT) that reflect the

meanings of lexical units in the source text and semantic dependencies among them, performed at the sentence level but often having to take into account suprasentential contexts. Semantic analysis procedures are typically developed for a particular domain (e.g. medicine, finance, and computers), though general, "common sense" semantic knowledge is also used. The existence of canonical formalisms for encoding world knowledge and text meaning enables the use of a single universal semantic interpreter with different knowledge source for each domain.

o Pragmatic or discourse analysis: suprasentential analysis leading to the resolution of anaphors, ellided phrases, deixis, as well as the attribution of intent and speech acts. In its full form, discourse analysis leads to the creation of a text-meaning structure in a representation language with the various domainoriented and rhetorical relations among the elements of a text, including coreference of noun phrases and anaphors, causal and temporal relations, topic/comment structure and so forth. The state of the art in pragmatic and discourse analysis is not as well developed as the other three phases of language analysis.[2,11]

Approaches to Machine Translation :[2]

Traditional approaches

Traditionally approaches of MT can be classified by their architectures:

o Direct or transformer architecture
o Transfer based architecture
o Interlingual architecture

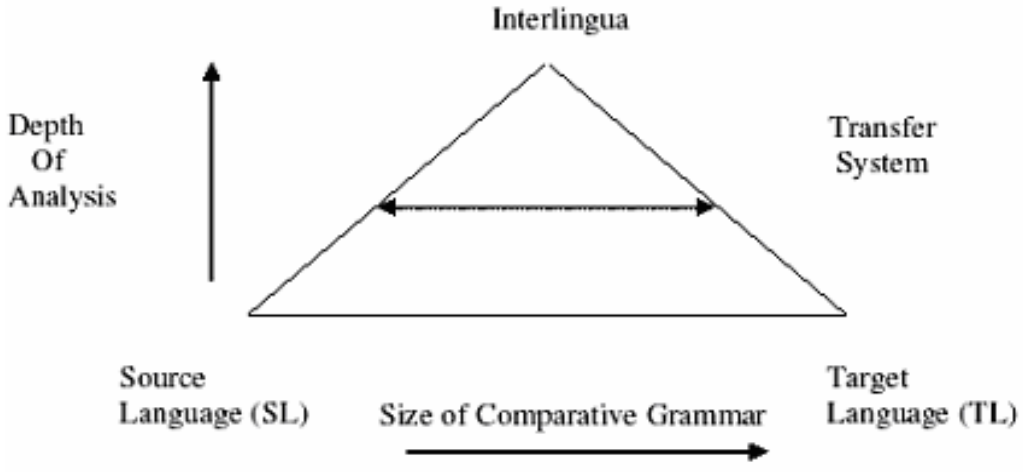

Fig (2) Traditional approaches of machine translation

Other approaches

A selection of approaches to MT that illustrate a range of useful techniques is presented.
Other approaches of MT can be classified as follow:

- Knowledge-based approach
- Corpus-based approach
- Hybrid approach

What are supported requirements to construct Machine Translation systems ? [5]

Source Language Dictionary (SL)
(SL-TL) Dictionary
Target Language Dictionary (TL)

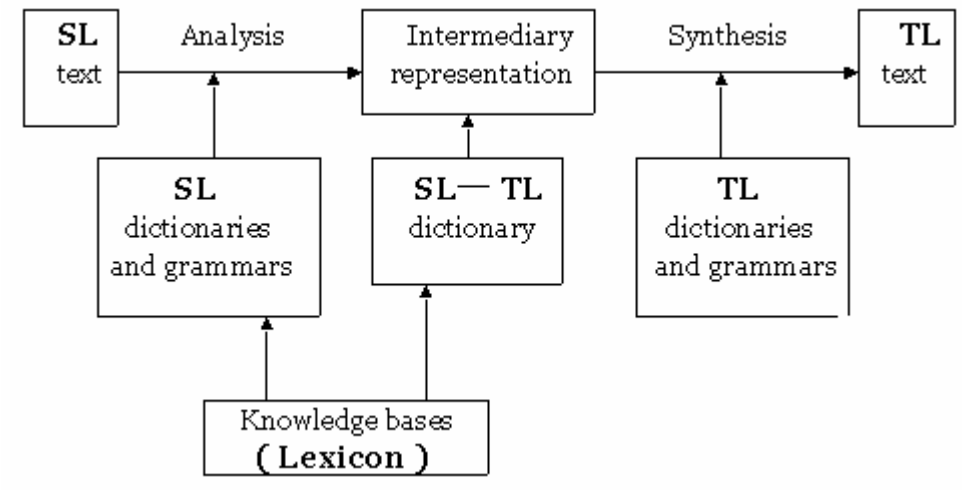

Fig (3) Machine Translator mechanism

Research methodology

What is the best method for dictionary organization (representation)?
It can be organized by one of the following: - [5, 6, 7]

1) **List of vocabulary**: showing words and their meanings (two languages) and adding other synonyms and derivations leading to fully correct translation.

2) **Relations Array :** English morphological relationship: This is a multi-dimensional array. This depends on the complexity of the dictionary and its standard for translation.

Arabic grammatical array : It is also multi-dimensional; its first dimension is the root of the word and the second is the meaning matching the root, and its other dimension represents its semantic significance and meaning. As an example for tri – array :

Ta'qdum qa'ddam - Taffa'al (motion)
Iftikhar Fakhir - Iftia'l (Action)

This is already suggested by this paper to represent knowledge arrays in the form of two-dimensional associational memory. The link would be done on different levels, each represents input for the level that precedes and follows according to the associational modes when a word to be translated is inserted.

3) **Database :** This is done by connecting files with each others to meet the needs the requirements of an effective database having no repetition or dependence. This would provide quick and accurate search. The database might be as follows:

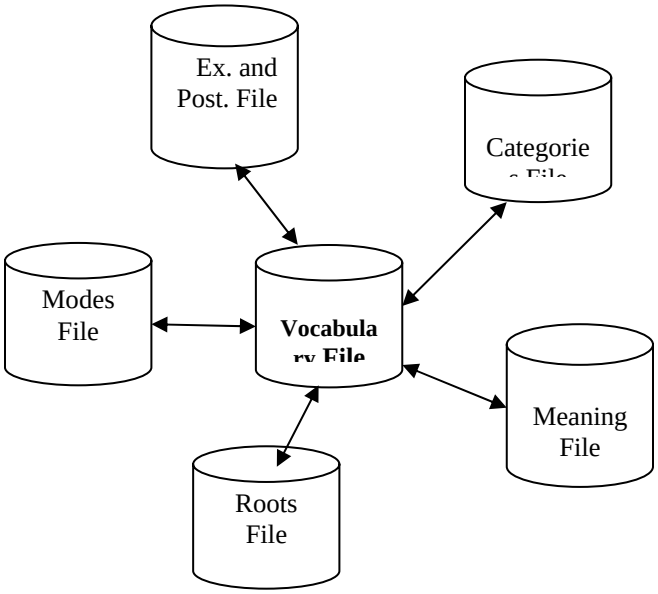

Fig (4) Database for dictionaries representation

Bi-Associative Memory

It is a memory used in connecting and recalling data to create feature almost similar to man's remembering process. For example, upon seeing a man approaching, one tries to identify such figure by recalling to modes of differentiation (like appearance). If congruity between such person and the data stored in memory is made , then the second process of recalling starts (like his name, age, voice, job, kinship, color, length, .. etc). This is the

associative memory process which connects a key with data through specific functions that make the recalling processes possible. [5, 8, 9, 10]

On the other hand, the Bi-Associative memory consists of two layers frequently repeated, the first one represents the input (X) and the second the output (Y) and vice versa i.e. ($X \leftrightarrow Y$).

Vector **X** having n size and vector **Y** having m size. Both vectors are connected by a group of conductors based on using **weight** matrix for connecting and recalling. This array is referred to by (W_{m*n}). The connection between **X**, **Y** layers are bi-directional which means that data is transferred from **X** vector to **Y** and from **Y** to **X**.

The weight matrix is used to store the juxtaposed modes, for example, weight (W_{ij}) represents connection between the mode arrangement from **X** vector and that from vector **Y** which helps the signed to be transferred in both directions.

The weight matrix (W) is as follows:

$$W = \sum W_p \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Which means the total product of the different associative modes. Each mode value is gained as follows:

$$W_p = X_p^T * Y_p \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

The value of each mode can be recalled by input it in any of the two directions and by considering the other mode through the following equation:

$$net\ Y(t) = W_{ij} * X(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

W_{ij} stands for the line value i^{th} from weight matrix. On the other hand, the opposite direction would be re-measured by Transpose Matrix for the X^T. This process continues till stabilization of the network is achieved i.e. giving the same value for the two successive inputs.

It was noticed that Bi-Polar arrays are more efficient and accurate compared to ordinary arrays [Kosko 1987].

The Bi-Polar array is done as follows:

$$X_i(t+1) = \begin{cases} +1 & net\ X_i(t) > 0 \\ X_i & net\ X_i(t) = 0 \\ -1 & net\ X_i(t) < 0 \end{cases} \quad \ldots\ldots\ldots\ldots \quad (4)$$

As $net\ X(t) = W_{ij} * Y(t)$, For illustration, lets see the following example:

Suppose you like to connect the following modes:

$X_0 = (-1 \quad +1 \quad +1 \quad -1 \quad +1) \leftrightarrow Y_0 = (+1 \quad +1 \quad +1 \quad +1)$

$X_1 = (+1 \quad +1 \quad -1 \quad -1 \quad +1) \leftrightarrow Y_1 = (+1 \quad +1 \quad -1 \quad -1)$

$X_2 = (+1 \quad -1 \quad +1 \quad -1 \quad +1) \leftrightarrow Y_2 = (+1 \quad -1 \quad -1 \quad +1)$

We must find the W_p resulting from equation (2):

$W_0 = X^T_0 * Y_0$

$W_1 = X^T_1 * Y_1$
$W_2 = X^T_2 * Y_2$

Then we find the weight matrix (W) from equation (1):

$$W = \begin{bmatrix} +1 & -1 & -3 & -1 \\ +1 & +3 & +1 & -1 \\ +1 & -1 & +1 & +3 \\ -3 & -1 & +1 & -1 \\ +3 & +1 & -1 & +1 \end{bmatrix}$$

Let us suppose that what is required is to recall the mode accompanying X_2 , then first we must fined ***net Y*** from equation (3): ***net Y= (+7, -3, -5, +5).*** By equation (4) to find the Bi-polar vector, we shall have ***net Y = (+1, -1, -1, +1)*** , which stands for the accompanying vector Y_2 for X_2 vector.

It is noteworthy that the process of recalling the mode cannot always be done easily. Sometimes the process fail due to the initial generally process of the tow vectors ***Y, X***. This can be overcome by work systems and to rely on vectors which give correct values.

The New Idea

Arrays required for translation are bi-poles. Transformation of texts, data and programs into digital files are not difficult; it is done by ASCII code or by scanner or by sound processing, or by Forier transformations, or by using the physical properties. So, digital arrays for words processing and their morphology can be obtained and transformed into bi-poles arrays to be ready during the start of the recalling process.

Weight array for each level is to measured once following the same previous steps to recall the attached mode which represents the required word. The word is to be inserted after omitting affixes if any. Then root of the word would be recalled to represent input level that follows, then the syntactical morphology as new input through which it attached mode would be recalled representing the mould of the word and its meaning as follows :

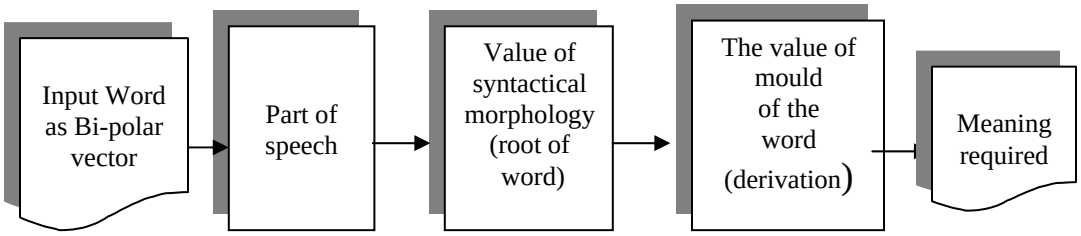

Fig (5) Recalling method

The above figure can applied to recall the meaning of the input word. See the following example :

English Word	**Part of speech**	**Root**	Derivation	**Meaning**
Friend	n.	Sidg	Fa'eel	Sadeeq
Friendship	n.	Sidg	Fa'eel	Sadaqa
Befriend	v.	Sidg	Yufael	Yusadiq

Conclusion and Future Work

Recalling process of data is an important process in translation from English into Arabic. In This paper, BAM neural networks would be used to achieve two-different recalling mode. This requires appropriate analysis of the linguistic potentials of each word to obtain accurate data that would provide appropriate architecture for the bi-direction associated memory.

This design would provide efficient way for performance that help minimize storage capacity by storing modes of sizes less than those required by normal storage way. This means that a word of 8 bit can be stored as attached mode at 6-bit size or less. This also provides wide capacity of storage in case of large number of vocabulary and their meaning. Adding new vocabulary does not affect the recalling process as in the traditional ways. Besides, both ways of treatment are made one. The easiness of treatment is shown in word processing whether those odd ones or ordinary as it follows a fixed base as the dictionary deals with the vector of the words and not with words themselves.

- This method would be used in the electronic dictionaries –double languages (English-Arabic/Arabic-English) whereby the vector could be measured from any given direction.
- Analyzing and understanding texts.
- Designing interfaces for natural languages systems.
- Machine Translation.
- Machine Learning.
- Computerized Learning.
- Accepting a written short story text to be transformed into a visual animation.

References

1- Chomisky (Muard Azez :translator), **"Syntactical Structuers" ,** A hundred book serious, General Thompson Culture publishing, Iraq, 1987.

2- Mohamed, Azza A., "**Machine Translation of noun phrases :from English to Arabic** ",M.Sc. thesis, University of Cairo, Egypt, 2000.

3- Al Najjar, Majed Flyeh, Murad, Mohammed N., **"English into Arabic machine translation: syntactical lexicon transformations between the two languages"** , Iraq, 1996.

4- Al-Gaphari, Ghaleb H. "**A Constraint- Based Object Lexicon for Supporting Natural Language Processing**", Ph.D. thesis, University of Basra, Iraq, 1999.

5- Batiha, K; Yousif, J.H.,"**The representation of Lexicon Using BAM Supporting MT"**, Zerqaa University 2001.

6- Patterson, Dan. W., **"Artificial Neural Networks"**, Prentice Hall, 1996.

7- Duamy,Mehdy gizar. **"A vocabulary Design to use fin machine translation system"**, M.Sc. Thesis, University of Saddam, Iraq, 1992.

8- Fadhel, muaed abdul razaq, **"A Lexicon Design to support English into Arabic Machine translation"**, Ph.D. thesis, University of Technology, Iraq, 1977.

9- Cynthia, A.; Mooney, Raymond J., "**Acquiring Word-Meaning Mapping for Natural Language Interfaces"** , submitted to Journal of Artificial Intelligence Research, 2002.

10- Firebaugh, Morris W. "**Artificial Intelligence A Knowledge Based Approach**", PWS-Kent Publishing Company, Boston, 1988.

11- Athman, Zainab A., "**Machine Translation from Arabic to English for some verbal sentences in arabic** ",Ph.D. thesis, University of Basra, Iraq, 2007.