Understanding Natural Language through the UNL Grammar Workbench

Sameh Alansary
Bibliotheca Alexandrina, Alexandria, Egypt
&
University of Alexandria, Faculty of Arts, Department of Phonetics and Linguistics, Alexandria, Egypt.
Sameh.alansary@bibalex.org

Magdy Nagi
Bibliotheca Alexandrina, Alexandria, Egypt
&
University of Alexandria, Faculty of Engineering, Dept. of Computer and System Engineering, Alexandria, Egypt
Magdy.nagi@bibalex.org

Noha Adly
Bibliotheca Alexandrina, Alexandria, Egypt
&
University of Alexandria, Faculty of Engineering, Dept. of Computer and System Engineering, Alexandria, Egypt.
Noha.adly@bibalex.org

Abstract

This paper is an attempt to introduce the Universal Networking Language (UNL) as a tool for achieving natural language understanding through a capable grammatical framework. The UNL system utilizes a processing workbench capable of effectively and accurately extracting the universal meaning behind the sentences of any language and, thus, analyzing and generating natural language words, sentences and texts. Such a framework can subsequently be used by linguists in the field of natural language processing (NLP) for applications such as machine translation (MT), question answering, information retrieval, etc.

1 Introduction

The field of natural language processing was at some point in time referred to as Natural Language Understanding (NLU). However, today, it is well agreed that NLU represents the ultimate goal of NLP. Yet, that goal has not yet been accomplished as a full NLU System should be able to: a) paraphrase an input text; b) translate the text into another language; c) answer questions about the contents of the text; d) draw inferences from the text (Liddy, 2001)

Thus, the UNL system attempts to fulfill all of the previous criteria by meticulously and accurately analyzing the input text into a universal abstraction of meaning. This meaning is represented in the form of a semantic network (the UNL network, or UNL expression). In this network, concepts are represented language-independently in the form of nodes, and each node is augmented with a wealth of semantic, grammatical and pragmatic information.

The grammatical foundation of the UNL system, thus, draws upon this information to determine the pure semantic relation between nodes. By determining them, the UNL can be said to have understood the natural language sentence; it can paraphrase this network into the same or other language, it can deduce certain information from its contents, etc. Moreover, using its robust grammars, the UNL system can generate a new meaning altogether and then generate it as a natural language sentence, in any language chosen.

The UNL grammar framework mainly adopts the X-bar theory as a foundation. The X-bar theory is in many respects similar to the UNL approach to natural language understanding. It assumes binary relations between the sentence constituents, which facilitates the process of mapping syntax onto semantics and vice versa. The X-bar theory also allows for many intermediate levels, a fact that gives the UNL great flexibility in the formation and decomposition of deep and surface syntactic structures.

In this paper, section 2 will start by examining the process of analyzing a natural language sentence. The process involves determining the exact meaning of words and the abstract relations they hold together in addition to encoding the other semantic, grammatical and pragmatic information they carry. In section 3, on the other

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

hand, the process of natural language generation is discussed. The capabilities of the generation grammar become clear in the way it is able to generate the constituent words in the target language, arrange them grammatically and make the necessary changes to the form of the word.

It is worth noting here that the arrangement of the following sections does not reflect the workflow of the UNL system or the ordered stages through which a natural language passes until it is finally transformed into a UNL semantic network, or vice versa. All of the following processes whether in analysis or generation work in unison and simultaneously to reach the most accurate understanding possible.

2 Analyzing Language

In order to claim the ability to fully and accurately understand human languages, a system must have the tools and methods capable of effectively decomposing the sentence into its basic constituents, understanding and encoding in some formal manner the intended meaning behind each constituent and the meaning reflected by its superficial grammatical form as well as its position in the sentence. It should also understand and encode the semantic relation between each constituent and the others.

The following subsections will present the techniques adopted by the UNL system to carry out the above processes; first, how a word in a natural language sentence is decomposed, analyzed and assigned the labels capable of bringing about a coherent and consistent sentence; second, how these words are linked together to form a syntactic structure then a semantic network that reflects the meaning of the whole sentence.

2.1 Analyzing Words

Words are the main conveyors of meaning in a sentence. The morphemes constructing a sentence carry the information without which a natural language sentence would be incomprehensible. "A positive absolute universal is that the morphemes of every language are divided into two subsystems, the open-class, or lexical, and the closed-class, or grammatical (see Talmy, 2000a). Open classes commonly include the roots of nouns, verbs, adjectives and adverbs and contribute most of the content. Closed classes, on the other hand, determine most of the structure and have relatively few members. They include bound forms such as inflections, derivations, and clitics; and such free forms as prepositions, conjunctions, and determiners (Talmy, 2000) (Francis, 2005).

Encoding Functional and Grammatical Morphemes

To understand the full meaning of a sentence, closed classes must be acknowledged as they contribute to the meaning by cross-referencing the main concepts and indicating other linguistic and extra-linguistic information. Due to the semantic constraints on the closed-class subsystem, they constitute an approximately limited inventory from which each language draws in a unique pattern to represent its particular set of grammatically expressed meanings (Francis, 2005). This inventory is mimicked in the UNL system by a set of tags capable of representing the grammatical, semantic, pragmatic information that might be conveyed by the closed-class morphemes in any language (Alansary et al., 2010)[1].

Closed classes may be either represented as bound morphemes or free morphemes; bound morphemes are usually the result of inflection or derivation processes. The Arabic language, for example, is highly inflectional and is especially rich in word forms. Thus, Arabic word such as فتجاهلوني fatagaahaluunii 'So they ignored me', although a single orthographic word in Arabic, it is the equivalent of a whole phrase in some other languages. Therefore, in order to understand the full meaning of such a complex word, the information communicated by the bound morphemes in it must be included into its meaning.

Uncovering the bound morphemes in a word (i.e. affixes) and what they represent involves separating them from the core open-class concept by scanning the input words and matching them with the entries in the natural language-UNL dictionary; the longest most appropriate string match is chosen. However, there are usually several matches and, consequently, several potential analyses for a single input word. For example, figures 1 and 2 show two of the potential morphological analyses for the previous example word فتجاهلوني.

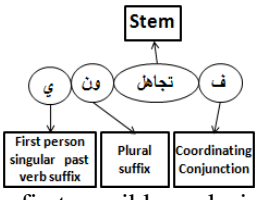

Figure 1. The first possible analysis for فتجاهلوني

[1] This set of tags and information about each is available at http://www.unlweb.net/wiki/index.php/Attributes

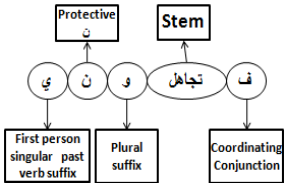

Figure 2. The second possible analysis for تجاهلوني

To resolve this sort of ambiguity, morphological disambiguation rules are used. Disambiguation rules assign priorities to the possible sequences of morphemes. Priorities range from 0 (impossible) to 255 (possible). In this case, the rules in (1) apply.

(1a) (^PRS,^PREFIX)(VER)("ون"):=٠;
(1b) (V,PAST)("و",VSUFFIX,SPRON):=255;

Rule (1a) disqualifies the first analysis (figure 1) by stating that a past verb (not preceded by any of the present tense prefixes) can never have a "ون" as a suffix. On the other hand, rule (1b) maximally approves the string "و" being a suffix for a past verb. In the same manner, all of the constituent morphemes are disambiguated and the wrong analyses are refuted until only the most appropriate analysis is left which is the analysis in (figure 2).

In addition to bound morphemes, other closed-class members are free forms such as conjunctions, prepositions, auxiliaries, etc. These are also scanned and matched with dictionary entries; however, these morphemes are encoded using special tags into the final semantic network[2]. For example, conjunctions such as بعد baEd 'after' and إذا ?ithaa 'if' are replaced by the tags "@after" and "@if" respectively. While adpositons like فوق fawq 'above' and حتى ħattaa 'until' are replaced by "@above" and "@until" respectively.

Encoding Main Concepts

Aside from the previous grammatical or functional morphemes, the main content conveyed by a word is carried by a nominal, verbal, adjectival or adverbial lexical item that belongs to the open-class. After abstracting away all the functional bound morphemes from a word, the stem representing the main concept is left behind.

It is claimed that any of the concepts a person can know ought to have the potential to be expressed in any human language and that the semantic representations of words would be a particular type of concept (Francis, 2005). Thus, the UNL system has taken up a sort of language-independent conceptual representation to replace the members of the open-class words. This representation is a machine-readable numerical ID that stands for the exact sense the natural language word usually means. This ID is, in fact, adopted form the English WordNet 3.0. In the WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct unique sense (Fellbaum, 1998). The UNL system, then, uses the ID given to each synset in the WordNet ontology to refer to the universal concept it denotes. The use of the WordNet in building the UNL dictionaries and ontologies is discussed in (Boguslavsky et al., 2008) (Bekios et al., 2007), (Martins and Avetisyan, 2009), (Boudhh and Bhattacharyya, 2009).

For example, when encountering the Arabic word فتجاهلوني, the detected stem تجاهل tagaahala will be matched with the entries in the main natural language-UNL dictionary to determine the universal concept it stands for. However, this stem is also prone to some sort of ambiguity; it can either represent a noun meaning 'ignoring' or a verb meaning 'ignore'. To determine which interpretation is intended here, lexical disambiguation rules come into effect; the rule in (2) resolves this ambiguity.

(2) (NOU,MASDR)("و"|"ون"):=0;

This rule rules out the possibility of "تجاهل" being a noun since the possible noun alternative is a verbal noun and a verbal Arabic noun can never have a "و" or a "ون" as suffixes. Thus, "تجاهل" is established as a verb. However, even as a verb, there are five alternative senses to choose from[3]. Nonetheless, it can be argued at this point that the word is indeed understood in the sense that only the most appropriate interpretations are listed, all of which would be correct in different contexts and under different circumstances. For the purposes of our discussion, it will be presumed that this word is equivalent to the

[2] These tags representing these too are found at http://www.unlweb.net/wiki/index.php/Attributes

[3] The process of choosing the exact ID to represent the intended sense of the natural language lexical item is not an easy process. Apart from grammar rules, the UNL system makes use of extensive ontologies and knowledge bases that aid the process of word-sense disambiguation. However, this paper will only focus on the grammar-related solutions to word-sense disambiguation; the other techniques will be thoroughly discussed in forthcoming publications.

universal representation "200616857" which means "give little or no attention to".

Encoding Semantic, Grammatical and Pragmatic Information

Finally, after representing all the previous forms of orthographically-represented information (bound and free morphemes), other subtle information are also extracted and are meticulously included in the understanding of words. This subtle information is in fact included along with the definitions of concepts in the natural language-UNL dictionary. They include semantic, grammatical and pragmatic features carefully selected from the tagset the UNL system employs. Semantic information such as abstractness, alienability, animacy, semantic typology (cognitive verb, communication verb, location, natural phenomena, etc.) and others are included with the entries representing each of the constituent concepts of a sentence. Figure 3 illustrates some of the semantic information assigned to the entry representing the concept "103906997" meaning "a writing implement with a point from which ink flows"; i.e., قلم حبر 'pen'.

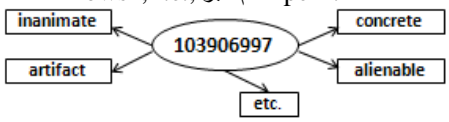

Figure 3. Some of the semantic information included with the concept for the Arabic lexical item قلم حبر in the Arabic-UNL dictionary

Moreover, entries in the natural language-UNL dictionary are also ascribed grammatical information. Grammatical information includes the concept's lexical category, lexical structure, part of speech, transitivity, valency, case, syntactic role, etc. Figure 4 shows some of the grammatical features assigned to the conceptual representation "202317094" meaning "bestow, especially officially"; i.e. منح 'grant'

Figure 4. Some of the grammatical information included with the entry for the Arabic word منح in the Arabic-UNL dictionary

In addition to the kinds of information acknowledged so far, there are also other pieces of information that are inseparable from any sort of understanding of a text. These types have to do with the pragmatic meaning of a sentence. It includes, for example, the situational context. The situational context would to the extent possible refer to every non-linguistic factor that affects the meaning of a phrase. The UNL, of course, cannot account for *all* of these factors but it can, however, detect and encode *some* of them. For example, an important element in the situational context is the role of a word in a sentence. Therefore, in the UNL framework, the main (starting) node of a semantic network is marked. Similarly, in passive constructions, the subject is indicated by a tag that specifies it as the *topic*.

A different sort of context markers are those indicating information on the external context of the utterance, i.e., non-verbal elements of communication, such as prosody, sentence and text structure, and speech acts. Linguistic context is also encoded; special tags denote the linguistic neighborhood of a word such as punctuation marks and anaphoric references.

Finally, a further type of extralinguistic information has to do with the social context of the sentence. When marked explicitly by the use of certain words or structures, information about the social context is also included in the understanding of a sentence showing, for example, social deixis (politeness, formality, intimacy, etc.) and register (archaism, dialect, slang, etc.) and others[4].The acknowledgment and inclusion of all such tags is quite necessary to claim the ability to truly understand a natural language sentence. Besides, they must be tagged in order to support later retrieval (Dey, 2001) as will be shown in the section 3 of this paper.

2.2 Analyzing Sentences

Understanding and encoding the meanings conveyed by every single morpheme in a certain sentence is far from sufficient to constitute an understanding. A simple list of concepts and tags will be hardly comprehensible even for the native speaker. Grammar rules are required to link these morphemes into a semantic network that represents the meaning of the sentence as a whole.

Deducing the pure semantic meaning directly from a simple list of concepts can be deemed impractical if not impossible; hence, the UNL system has opted for the use of an intermediary stage that maps this list onto an elaborate syntactic tree structure. The ordering of constituents in this tree and the syntactic relations that link them

[4] These tags can also be found at http://www.unlweb.net/wiki/index.php/Attribute

together can, subsequently, help point out the kind of semantic links the sentence implies.

After encoding the information carried by each lexical item in section 2.1, grammar rules use this information to carry out the process of determining the syntactic structure underlying the input sentence. The UNL grammar workbench is divided into two main categories: transformation grammar and disambiguation grammar. Transformation grammar comprises the rules capable of extracting the abstract meaning conveyed by the input sentence while disambiguation grammar involves lexically, syntactically and semantically disambiguating the natural language sentence in order to reach the most accurate UNL representation possible (Alansary et al., 2010). Both Transformation and disambiguation grammars involve several phases and types of rules[5]. Yet, this paper will not delve deeply into the details of these phases or types (they have been discussed before in Alansary et al., 2010; Alansary, 2010); this paper rather aims at demonstrating how these grammars are capable of handling and deciphering natural language phenomena.

Determining Syntactic Structure

A significant section of the UNL transformation grammar is devoted to transforming the incoming natural language list into an elaborate syntactic structure. To demonstrate this process, the Arabic sentence in (3) will be used as an example.

(3) منح الرئيس قلادة النيل لمجدي يعقوب

manaħa ʔarraʔiisu qiladata ʔanniili limagdii yaʕquub 'The president granted the Nile Medal to Magdi Yacoub'

The information assigned in the previous stage will come into use here; transformation grammar rules use the grammatical, semantic and pragmatic information as guides to determine the syntactic position each morpheme holds in the syntactic structure of the sentence. For example, the rules in (4) transform the natural language sentence in (3) into the syntactic tree in figure 5.

(4a) (V, %01)(N,HUM, %02):=VS(%01;%02);
(4b) (V,%x)(N,NONHUM,%y):=VC(V,%x;N,%y);
(4c) (V,TST2,%01)PP("ل";%02):=VC(%01;%02);

Rule (4a) specifies that when a verb is assigned the semantic feature "give verb" and is followed by a "human" noun, the noun is the syntactic specifier of the verb. Rule (4b), on the other hand, states that the syntactic relation between a "give verb" and a following "non-human" noun is a syntactic complementizer relation. Finally, a grammatical feature of the verb "منح"; it being "ditransitive", dictates that when being followed by a prepositional phrase headed by the preposition "ل", the prepositional phrase is a second complementizer for that verb.

Along with these transformation rules, disambiguation rules are at work. Tree disambiguation rules also prevent wrong lexical choices and provoke best matches by determining which constituents can share a syntactic relation. For example, the rules in (5) help restrict the application of transformation rules by dictating that a prepositional complementizer (PC) following a distransitive verb (TST2) can never be an adjunct for that verb (probablility = 0) while it being a complementizer for that verb is very plausible (probability = 225) since a ditransitive verb inevitably requires two complements

(5a) VA(TST2,PC):=0;
(5b) VC(TST2,PC):=255;

The result of these processes would be the syntactic structure in figure 5.

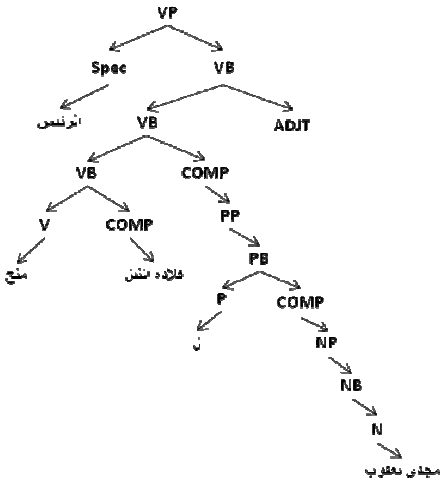

Figure 5. The deep syntactic structure of the Arabic sentence in (3)

Determining Semantic Structure

Finally, in order to generate the final understanding of an input sentence, different types of trans-

[5] More information about this division and the phases involved is found in http://www.unlweb.net/wiki/index.php/Grammar_Specs

formation rules apply to transform the syntactic structure in figure 5 into a semantic network. This sematic network will incorporate all of the information extracted and defined in the previous stages. Using the same example sentence in (3), the functioning of the semantic analysis rules will be demonstrated. The rules in (6) use the syntactic relations as a guide to determine the semantic links between the concepts.

(6a) VS(%01;%02):=agt(VER,%01;%02,NOU);
(6b) VC(%01;%02):=obj(VER,%01;NOU,%02);
(6c) VC(%01;PC("٠٢%;"ل)):=gol(VER,%01;NOU,%02);

The rule in (6a) states that if the specifier of a verb is a noun syntactically, then the noun is the agent of the verb semantically while rule (6b) assigns a semantic object relation between two words that share a syntactic complementizer relation. In the same manner, rule (6c) states that if the complement of a verb is a noun syntactically and it is a prepositional phrase introduced by the preposition (ل), then the noun is the goal of the verb semantically.

Also on the semantic level, disambiguation rules (i.e. network disambiguation rules) apply over the network structure of UNL graphs to constrain the application of transformation rules. Disambiguation rules constrain the type of constituents to be a member in a binary semantic relation. However, in this example sentence, no network disambiguation rules were required.

All of the previous processes work in unison to finally generate the semantic network in figure 6.

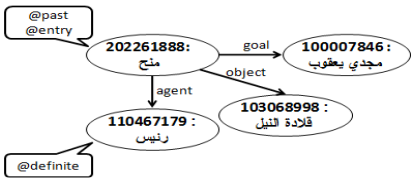

Figure 6. The semantic network representing the the Arabic sentence in (3)[6]

3 Generating Language

A capable understanding framework is not only essential to the process of analyzing and processing natural language, Natural language Generation (NLG) can be deemed impossible if an adequate understanding foundation is unavailable. An efficient NLG system must be able to generate language accurately in order to answer questions, for example, or interact with the user for the purposes of translation, information retrieval, etc.

In the following subsections, the UNL grammar workbench as an efficient and robust means for generating language will be considered. The process of generation may be seen to some extent as the mirror image of the analysis process; the abstract meaning stored inside the machine in the form of an extensive semantic network is transformed into a natural language sentence in two main stages. First, the whole structure of the sentence to be generated is determined on the deep level then on the surface level. Second, the word forms necessary to convey that meaning are generated to fill in the slots in this structure and the final changes are made to form a comprehensible well-formed natural language sentence.

Similar to the process of analyzing natural language, generating well-formed sentences has to pass through five stages of transformation rules, in addition to disambiguation rules; passing from the abstract semantic network to a syntactic representation from which the final natural language sentence is generated. This arrangement of phases is not the main focus here; it is rather to demonstrate the types of rules at work and how they are able to generate language from meaning efficiently as will be shown in the following subsections.

3.1 Generating Sentences

A syntactic structure is indispensible to constitute a well-formed target language structure. Thus, the UNL framework uses a set of formal rules to translate the pure semantic links that make up the abstract meaning representation (i.e., the UNL network) into syntactic relations.

Generating Syntactic Structure

There are two types of syntactic structure; the deep structure and the surface structure. The deep structure of a sentence represents its meaning but interpreted using syntactic tags rather than semantic ones. The surface structure, on the other hand, reflects the ordering of the constituents in the final natural language sentence.

In the process of forming a sentence's deep structure, grammar rules are devoted to mapping the semantic relations from the semantic network onto their equivalents in a syntactic tree. As an example, the semantic network in figure 6 requires the mapping rules in (7) to map the se-

[6] This semantic network only represents the main structure of the sentence in (3) and some of the most crucial tags; it does not incorporate all of the semantic, grammatical, and pragmatic information detected because of the space limitation.

mantic agent, object and goal relations onto their counterpart syntactic relations: verb specifier, verb complementizer and second verb complementizer, respectively.

(7a) agt(VER,%01;%02,NOU):=VS(%01;%02);
(7b) obj(VER,%01;NOU,%02):=VC(%01;%02);
(7c) gol(VER,%01;NOU,%02):=VC(%01;PC("٠٢%;"ل));

The mapping rule in (7a) states that if the agent of a verb is a noun semantically, then the noun is the specifier of the verb syntactically and thus, occupies the specified positions in the syntactic tree. Similarly, the rule in (7b) maps the semantic object relation onto the position of a complementizer relation syntactically. Finally, rule (7c) maps the semantic goal relation onto the position of a (second) complementizer relation. The result, of course, would be the same syntactic structure in figure 5 above.

However, this deep structure does not always reflect the correct ordering of constituents in the final natural language sentence. The constituents of the tree have to be mapped onto a morphological sequence that is considered well-formed according to the grammar of the target language. This ordering is determined in the surface syntactic structure; thus, this deep syntactic structure has to be mapped onto a surface structure before being generated as a natural language sentence.

A sentence may have multiple surface structures since the same meaning may be reflected in several synonymous sentences. The Arabic language is especially abundant in such cases because Arabic word order is comparatively free; although the canonical order of an Arabic sentence is VSO (Verb-Subject-Object), most other orders can occur under appropriate circumstances (Ramsay and Mansour, 2006).

Thus, a different type of grammar rules is subsequently used to determine the exact position of a constituent with regards to the others, when certain conditions are fulfilled. For example, the rules (8) and (9) can generate two different equivalent versions of the syntactic structure in figure 5; these two versions are shown in figures 7 and 8, respectively.

(8)VB(VB(%x;%y);%z)VS(%x;%v):=VP(VB(VB(%x;%v);%y);%z);
(9)VB(VB(%x;%y);%z)VS(%x;%v):=VP(VB(VB(%x;%v);%z);%y);

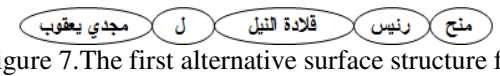

Figure 7.The first alternative surface structure for the deep structure in figure 5

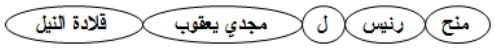

Figure 8. The second alternative surface structure for the deep structure in figure 5

The rule in (8) positions the second complement "مجدي يعقوب" before the first complement "قلادة النيل" to generate the sentence in figure 6. On the other hand, the rule in (9) positions them reversely to generate the sentence in figure 8.

Rules such as the previous apply generally to position constituents that behave regularly; nonetheless, in case of exceptions or categories that do not follow a regular distributional pattern the ordering of the constituent is informed in the dictionary. For example, the Arabic preposition ل 'to' is assigned the tag "IEMT" to indicate that, unlike most other prepositions, this preposition immediately precedes the noun following it, without and intervening blank space. This is indicated by the distribution rule in (10).

(10) PB(PRE):=PB(+IBEF)

Moreover, in special cases, the ordering specified in the surface structures needs to undergo some adjustment to reflect some aspect of meaning. In such cases, movement rules rearrange the constituents in a sentence to deal with transformations that affect only the surface structure of the sentence such as topicalization and passivization. For example, the movement rule in (11) changes active structures headed by monotransitive verbs into passive by changing the position of a the verb complementizer to fill the place of the verb specifier, while the verb specifier moves into the position of the verb adjunct as a prepositional phrases headed by the preposition بواسطة.

(11)VC(%head;%comp)VS(%head;%spec):=VS(%head;%comp)VA(%head;PC([بواسطة];%spec));

Generating Functional Morphemes

Up to this point in discussion, the natural language sentence is still an abstract list of concepts. As mentioned earlier, the semantic network is not only composed of nodes representing the main concepts and the semantic relations that tie these concepts together. Each node in this network is assigned numerous tags that signify the omitted closed-class free forms such as particles, prepositions, conjunctions, auxiliaries, interjections, quantifiers and others.

Closed-classes must also be acknowledged in the deep and surface structures of a sentence.

Tsherefore, parallel to the previous section, other types of rules are at work to express these closed-classes in the form of free morphemes, and position them in the syntactic structures being formed. For example, the rule in (١2) generates and positions the Arabic definite article "ال" in a surface structure such as the one in figure 7 as illustrated in figure 9.

(12) @def := NS(DP([ال]));

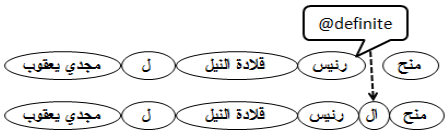

Figure 9. The surface struture in figure 7 after generating and position the definite artice "ال"

Moreover, in some cases and in some languages a grammatical feature has to be expressed independently as a free morpheme; a phenomenon called periphrasis. An example of this phenomenon is the present perfect tense in Arabic which is formed by adding the particle قد qad before the present verb. The rule in (١٣) generates this construction.

(13) VH(%vh,PRS,PFC):=+IC([قد];%vh,+PAS);

This rules states that the head of the verbal phrase receives the feature PTP (past simple) and becomes the complement of an inflectional phrase headed by the lemma قد if it has the features PRS and PFC (present and perfect).

In addition to grammatical and functional morphemes, some semantic relationships are expressed in some languages subtly through word order, inflection or derivation, while in other languages some relation has to be expressed in the form of free morphemes. Consequently, when generating Arabic sentences, some of the semantic relations used within the UNL framework had to be included in the syntactic structures as distinct Arabic lexical items. An example is the UNL semantic relation "material" which has to be expressed in Arabic as مصنوع من maSnuuEun min 'made of' to link between the object and the material. This is illustrated in the sentence قميص مصنوع من القطن qamiiSun maSnuuEun min ?alquTni 'a cotton shirt' as shown in figure 10.

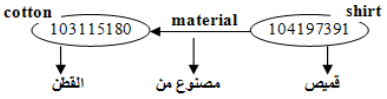

Figure 10. Generating the Arabic words مصنوع من to replace the semantic relation "material"

3.2 Generating Word Forms

In addition to the core concept conveyed via a natural language word, its superficial form in a sentence tells a lot about the role it plays in a sentence and the relationship it holds with other words. Moreover, incorrect word forms can deem a sentence incorrect and incomprehensible according to the grammatical regulations of the target language. Therefore, the final step to generating a well-formed natural language sentence is generating the required word form of each of the constituent lexical items. In this stage, grammatical information will be reflected on the form of the word itself, unlike the previous section where some grammatical or functional features had to be expressed as free morphemes.

Inflection

Inflection is the phenomenon of changing or modifying the form of a lexical item to indicate a certain grammatical feature it should convey. For example, verbs are inflected (or conjugated) to indicate their tense, aspect, etc. while nouns, adjectives and adverbs are inflected (or declined) to indicate their tense, mood, voice, etc. Inflection takes place mainly through the process of affixation; prefixation, infixation, suffixation or circumfixation.

The grammar workbench adopted in the UNL framework provides the means for generating the required morphological elements and attaching them to the intended concept. Inflection grammar is triggered by the tags put on the intended concept in the semantic network such as PLR (plural), FEM (feminine) or 2PS (second person).

Inflection is handled through inflectional paradigms in case of regular behavior, and inflectional rules in case of irregular behavior. These two may also work together in words that exhibit quasi-regular behavior. For example, a single inflectional paradigm (14) can handle the generation of the plural forms أواصر?awAsir 'relation' out of the singular آصرة ?asirah 'relation' and the plural أواني ?awaaniy 'pots' out of the singular آنية ?aniyyah 'pot'.

(14) PLR:=[1]>"١","وا"";

This paradigm inserts the string "وا" after the first letter and deletes the final letter in the word to generate the plural forms.

On the other hand, an inflectional rule handles irregular inflections, and, is thus, only applicable to a single lexical item. An example is the Arabic word امرأة ?imraah 'woman' of which the plural

is نساء nisaa' 'women'. This case is handled via the affixation rule in (15) which replaces the whole string امرأة with the string نساء.

(15) PLR:="نساء";

Arabic verbs are even more complex; a single verb may have over 60 distinct forms. However, the inflection of Arabic verbs is fairly regular and is, therefore, easily computated in the form of formal inflectional paradigms that can generate, for example, the forms of the Arabic verbs سأل 'ask' and بنى 'build' of which some are shown in (16) and (17), respectively.

(16) سأل- سألا - سئلا- يسألان- سئل- سألوا- سئلوا - يسألون- سألت- سئلت- سألن- يسألن- سألت- سئلت- تسأل- سل- أنتما سألتما

(17) بنى - يبني- بنيا – يبنيان – بنوا – يبنون – بنت – تبني – بنتا – تبنيان – بنين – يبنين – بنيت – تبني - ابن

Agreement

Affixation rules assume the responsibility of generating all the required word forms to fit the tags explicitly marked in the semantic network. Nonetheless, in many cases, other constituents will imitate the tagged ones in some grammatical feature; a phenomenon that is called agreement. Agreement (or concord) is a form of cross-reference between different parts of a sentence or phrase, it happens when a word changes form depending on the other words it relates to. Special UNL grammar rules are devoted to determining which constituents receive or assign a grammatical feature and under what circumstances. (18) shows a simple conditional agreement rule. This rule specifies that an adjunct receives the gender from the noun if adjective.

(18) NA(ADJ):=NA(+RGEN);

A different kind of agreement of special importance to the Arabic language is that case marking. Usually a language is said to have inflectional case only if nouns change their form to reflect their case. Case marking is the process of assigning grammatical case values to dependent nouns for the type of relationship they bear to their heads.

The type of inflection an Arabic noun undergoes greatly depends on its case. A case-marking rule such as the one in (19) determines an adjective to be inflected for plural by adding the suffix "ين" rather than "ون" when modifying a noun in the accusative case.

(19)(%x,M500,MCL,ACC):=(%x,-
M500,+FLX(MCL&ACC:=0□"ين";));

Spelling changes

Other word-level changes in the form of a word may not depend on its structure or syntactic role in a sentence but rather on its linguistic neighborhood. Examples of such changes are changes in the spelling of a word as a result of contraction, assimilation, elision, etc. or capitalization in the beginning of a sentence (in Germanic languages) or the use of punctuation marks.

These kinds of transformations are handled by linear rules. Linear rules apply transformations over ordered sequences of isolated words in the UNL framework. Linear rules replace, add or delete certain characters in a word according to the contiguous characters from other words. For example, the Arabic definite article "ال" when preceded by the preposition 'ل'; the first letter from the definite article is deleted and the preposition immediately adheres to the remaining character from the definite article with no intervening blank spaces. The rule in (20) performs this process.

(20) (%x,M90,DFN,LAM):=(%x,-
M90,+FLX(DFN&LAM="\>"ل;));

4 Scope and Limitations

The UNL system currently supports the processing of 17 different languages. The main resources necessary for their analysis and generation (dictionaries and grammars) are being built by the respective institutions scattered all over the world. Yet, the UNL system is flexible enough to support any other natural language once the necessary resources are built.

These processing capabilities cover the morphological, semantic, syntactic and phonetic aspects of natural language texts. However, the phonetic module is not yet activated but will be in the near future. Also, the syntactic module is currently devoted to handling the basic syntactic structures; other more complex structures are to be focused on in later stages of development.

Nevertheless, the UNL workbench does not claim it represents the 'full' meaning of a word, sentence or text using these modules since 'full' meaning, as mentioned earlier, may depend on an infinite list of factors such as: intention, world knowledge, past experiences, etc. Although these factors are mostly known for the human speaker/writer and listener/reader, such factors are too subtle and subjective for any attempt of systematic processing.

Moreover, it must also be clear that the UNL system only represents the most 'consensual' meaning attributed to words and phrases, other

equivocal meanings are quite complex for a machine to infer. Thus, much of the subtleties of poetry, metaphors, and other indirect communicative behaviors are beyond the current scope of the system; the UNL system mainly aims at conveying the direct communicative meaning as it constitutes the most part of day-to-day communications.

Users can establish the validity of the UNL workbench by using it to process natural language phenomena. This has already been done by dozens of computational linguists in the various UNL language centers who are, at the present moment, using the workbench to produce the necessary resources. The workbench has been found sufficient, flexible and representative of the phenomena exhibited by the natural languages being handled.

5 Conclusion

A system capable of understanding natural language sentences is of potentially unlimited uses in the field of natural language processing. As this paper aimed to demonstrate, the UNL framework provides natural language processing experts with a vast array of tools and mechanisms that would aid them in the endeavor of reaching a true, full and accurate understanding of a natural language sentence. The most obvious application of this system is, of course, machine translation where the UNL semantic representation functions as an interlingua; however, machine translation is definitely not the only use. A language-neutral representation of meaning as opposed to syntactic matching should be of great use is areas such as cross-lingual information retrieval. Also, by distinguishing between main concepts and other secondary constituents, this system can be used in text summarization or text extension. Another fundamental use would be to use the understanding of texts as the source encoding for webpages which, upon request, can be generated in the natural language the user chooses.

References

Alansary, Sameh, Magdy Nagi and Noha Adly. 2006. Generating Arabic Text: the Decoding Component in an Interlingual System for Man-Machine Communication in Natural Language. In *proceedings of the 6th International Conference on Language Engineering,* Cairo, Egypt.

Alansary, Sameh, Magdy Nagi, and Noha Adly. 2006. Processing Arabic Text Content: The Encoding Component in an Interlingual System for Man-Machine Communication in Natural Language". In *proceedings of the 6th International Conference on Language Engineering*, Cairo, Egypt.

Alansary, Sameh, Magdy Nagi, Noha Adly. 2010. UNL+3: The Gateway to a Fully Operational UNL System. In *Proceedings of Egypt 10th International Conference on Language Engineering*, Cairo, Egypt.

Alansary, Sameh. 2010. A Practical Application of the UNL+3 Program on the Arabic Language. In *Proceedings of the 10th International Conference on Language Engineering*, Cairo, Egypt

Bekios, Juan, Igor Boguslavsky, Jesús Cardeñosa and Carolina Gallardo. 2007. Using Wordnet for building an Interlingua Dictionary. In *proceedings of 5th International Conference on Information Research and Applications,* I.TECH. vol.1, pages 39-46, Varna, Bulgaria.

Boguslavsky, Igor, Jesús Cardeñosa and Carolina Gallardo. 2008. A Novel Approach to Creating Disambiguated Multilingual Dictionaries. *Applied Linguistics*, vol. 30: 70–92.

Boudhh, Sangharsh, and Pushpak Bhattacharyya. 2009. Unification of Universal Words Dictionaries using WordNet Ontology and Similarity Measures. In *proceedings of the 7th International Conference on Computer Science and Information Technologies*, CSIT 2009, Yerevan, Armenia.

Dey, Anind K. 2001. Understanding and Using Context. *Personal and Ubiquitous Computing Journal*, vol. 5 (1): 4-7.

Fellbaum, Christiane, editor.1998. WordNet: An electronic lexical database. MIT Press, Cambridge, MA.

Francis, Wendy. S. 2005. Bilingual semantic and conceptual representation. In J. F. Kroll and A. M. B. de Groot, editors, *Handbook of bilingualism: Psycholinguistic approaches*. Oxford University Press, New York, NY, pages 251-267

Liddy, Elizabeth D. 2001. Natural Language Processing. In *Encyclopedia of Library and Information Science*, 2nd Ed.: Marcel Decker, Inc., New York

Martins, Ronaldo and Vahan Avetisyan. 2009. Generative and Enumerative Lexicons in the UNL Framework. In p*roceedings of 7th International Conference on Computer Science and Information Technologies,* CSIT 2009, Yerevan, Armenia.

Ramsay, Allan M. and Hanady Mansour. 2006. Local constraints on Arabic word order. In *Proceedings of 5th International Conference on NLP, FinTAL 200)*, pages 447-457, Turku.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Vol. I: *Concept structuring systems.* MIT Press, Cambridge