Arabic Machine Translation: A Developmental Perspective

Ali Farghaly

Abstract— **The field of machine translation has been dominated in the last two decades by statistical and machine learning approaches. Recently, prominent computational linguists [1]-[2]-[3] have expressed misgivings about the exclusive reliance on machine learning approaches to the neglect of the contribution of the symbolic approaches. Furthermore, several machine learning researchers [4]-[5] have recently acknowledged that incorporating linguistic knowledge in their machine learning applications resulted in marked improvement in performance. In this paper, we begin with a brief review of the early attempts at developing machine translation systems and in particular the first English to Arabic machine translation system released in the early eighties. The first generation of machine translation systems followed the direct approach. Part II, is devoted to the rise of the transfer approach in machine translation with an example from the SYSTRAN Arabic to English machine translation system. Part III documents the successes of statistical machine translation systems using the examples of the Language Weaver Arabic-to-English machine translation system and the crowd sourcing Google system. We also talk about the AppTek hybrid approach to machine translation. Part IV concludes the paper.**

Index Terms— **Arabic, rule-based, machine translation, crowdsourcing - Arabic syntax, morphology, machine learning, hybrid systems.**

I. EARLY MACHINE TRANSLATION SYSTEMS

THE invention of the digital computer in the 1940s inspired scientists to think of using the unprecedented speed of the computer to translate texts from one language to another. So inspired, scientists started to take practical steps to realize the dream and vision of Descartes who wrote in 1629 about a mechanical process to convert one human language to another. In 1949, Warren Weaver, the pioneer of machine translation, wrote a memorandum to his colleagues making four proposals for machine translation systems that go beyond word for word translation. Warren realized that many words in language were ambiguous and he proposed in his memorandum to solve this problem by examining the immediate context of the ambiguous word [7]. He also drew attention to the analogy between the structure of the human brain and the "logical machine". He concluded that the machine translation problem is solvable. He also suggested using the cryptic methods that linguists used in the Second World War for deciphering the German secret code. These cryptographic methods relied heavily on frequencies of letters, combination of letters and letter patterns. He also believed that underlying the statistical regularities of languages, there is a logical and universal foundation which could represent an alternative to translate from one language to another.

At the same time with the beginning with the cold war in the 1940s, there was an urgent need for crude machine translation because the United States decided it was essential to scan and interpret every Russian communication coming out of the Soviet Union. However, there weren't enough translators to keep up with the huge volume of Russian books and papers published in the Soviet Bloc at that time. The urgent need to translate Russian into English coincided with the invention of computers. It was not surprising then, that developing Russian to English machine translation systems would be one of the first tasks these "miracle" machines were set to perform.

The first demonstration of the feasibility of fully automated machine translation took place in New York on January 7th, 1954. On that day, Georgetown University and IBM demonstrated the first non-numerical applications and capabilities of the "new" electronic brain by demonstrating a fully automated Russian English machine translation system. The system embraced the commonly held view that a language consisted of a lexicon and a finite set of rules that could generate an infinite set of sentences. Surprisingly, the first Russian to English machine translation system had only 250 words and 6 syntactic rules. This experiment raised high expectations that probably within five years machine translation systems would be readily available. The promise was to develop a system that does not require pre-editing of the input while produces a reliable translation of the input text in the target language that is clear, intelligible requiring only stylistic modifications. No details were given about the actual linguistic processing in the system. For example, no information about dictionary content and lookup procedures were given. No account of how the syntactic analysis of the Russian sentences was performed and how the target English structure was selected. However, there were some references to reversing the order of pairs of sentences by assigning rules to the lexical items involved. Later on a more detailed description of the system is presented in [8]. Garvin [8] gives more detailed description of the dictionary. For example, the dictionary entries were sometimes stems, endings or full words. Each entry is associated with three codes; the first code indicates which of the six syntactic rules would apply, the

Manuscript received 26 May 2010.

Ali Farghaly is a Professor in Computational Linguistics, Senior Member of Technical Staff, Text Group, Oracle USA, CA; Adjunct Professor of Arabic Linguistics, Monterey Institute of International Studies, Monterey, CA, USA. E-mail: afarghal@miis.edu

second code would determine which contextual information are needed to determine the target translation while the third code indicates whether words to be inverted or not. At the technical level, the system represents the first attempt at non-numerical programming which presents developers with many challenges. Developers had to deal with character coding of Russian and how dictionary entries to be stored, what lookup procedure would be followed and how the syntactic rules would be coded and executed.

The Georgetown-IBM experiment and similar other work at the time were significant. First, it was proven that the digital computer could perform non-mathematical tasks such as machine translation. The system took advantage of the speed of the computer compared to human translators. It was also shown that the computer surpassed humans in that it would never forget, could work 24/7 without getting tired, and would never ask for a raise or a vacation. Third, the system demonstrated the need to specify and describe linguistic structures at different levels such the lexical and syntactic levels was demonstrated. Fourth, ambiguity of language was understood to be a problem although it was underestimated.

However, as Hutchins [9] stated correctly, the period from 1956 to 1966 was: "a decade of high expectations and disillusions". The promise to deliver fully automated machine translation systems with no pre-editing and only stylistic post-editing with a 95% accuracy was never achieved. Serious research proved that language structure was much more complex than previously thought and that translators use huge amounts of linguistic, domain-specific, real world and common sense knowledge that was not considered relevant at the time. The ALPAC report [10] concluded that machine translation was not viable given the state of knowledge at the time. Consequently, funding for research in computational linguistics halted in the USA and did not resume until the mid and late seventies.

II. The First Arabic Machine Translation System

The first English to Arabic machine translation system was developed in the late seventies by Weidner Communications Inc. which was located in Provo, Utah. The system was developed following the Direct Method which aimed to produce fully automated Arabic translations of unlimited English source documents but did not limit the translation to a specific domain. There was no pre-editing module although it included a module for post-editing if desired. As in all other MT systems that adopted the direct method, it was designed for a specific pair of languages: English as the source language and Modern Standard Arabic as the target language. The system consisted of two main stages: analysis of the source language and generation of the target language. The analysis of English was oriented to enable the correct generation of target language expressions employing a large bilingual dictionary as well as a dictionary for idiomatic expressions. The vocabulary and syntax of English was not analyzed in depth and only to the extent required to generate Arabic equivalents. Thus the system was unidirectional and did not perform deep syntactic or semantic analysis of the source language.

The system was commercially utilized by Omnitrans of California Inc. which used it for the purpose of translating the Encyclopedia Britannica into Arabic. This project was not completed for lack of funding. The Sultanate of Oman also licensed the Weidner English to Arabic machine translation system and used it to translate official English documents into Arabic. The author attests, from his professional experience at Omnitrans of California, that it was possible to get relatively reasonable output of the system by manipulating dictionary entries targeting specific domains. However, development of the English to Arabic system stopped shortly after the company was acquired by foreign investors in 1984.

III. The Transfer Approach to Machine Translation

A. Problems with the Direct Approach to Machine Translation

The direct machine translation approach did not rely on deep linguistic analysis of the source language. It involved superficial manipulation of the word order of the source sentence to make it look more similar to the order of the target language. Accordingly, machine translation developers and researchers soon realized that the direct method could not deal with the complexity of natural language. For example, what was thought to be a simple swapping operation switching the subject and verb from an SVO to a VSO structure turned out to be very complex. The translation of an English sentence like 'John loves Mary' into Arabic involves switching the order of the subject John to come after the verb in Arabic to become 'يحب جون ماري'. However, when the subject of the English sentence becomes complex as in 'The tall man who was wearing a red tie and a white shirt and was speaking in an Italian accent with his guests, greeted us warmly', identifying the length and boundary of the subject requires deep parsing. The direct approach would not yield accurate results for complex sentences like this because it did not incorporate the required syntactic rules for parsing such sentences. There was also a need to develop the technology to efficiently perform deep parsing and to represent complex disambiguation rules. The transfer approach to machine translation provided significant contributions on two fronts: the syntactic description of language and a new technology for the representation and processing of deep syntactic parsing. We will begin below with the progress made in the seventies and eighties in linguistic theory.

B. Progress in Linguistic Theory

The second half of the twentieth century witnessed a paradigm shift in linguistics when Noam Chomsky [11] [12] challenged the well established theory of structural linguistics [13]. Chomsky redefined the goals of linguistic theory to account for native speakers intuitions about their language rather than simply investigating a corpus and finding regularities in that corpus. He also challenged the view held by structuralists that a child is born with a "tabula rasa" i.e. with no knowledge of language at all. Structural linguists believe that it is through listening, imitating, and repetition that a child acquires the language of his people. Chomsky showed that comparing the linguistic knowledge that a child internalizes with the fragments he is exposed to in his early linguistic experience points into a gap that need to be accounted for. The explanation that Chomsky offers is that a child is born with innate knowledge of "Language", i.e. although he is born without knowledge of any specific language, nevertheless, he knows what language is. In Chomsky's terms he is born with Universal Grammar (UG). For Chomsky, this is the only explanation for the uniformity and remarkable speed of language acquisition. Chomsky also challenged the structuralists' position that in order to write a description of a language, they must obtain a corpus of the language and perform a "discovery procedure" to deduce the generalizations underlying the language. He argues that a corpus of native speakers' utterances represents only the performance of the speakers of the language. Performance is usually affected by lapses of mind, change of plans, fatigue, and distractions .etc. It is not always a true reflection of native speakers' knowledge of their language. Speakers very often recognize the ungrammaticality of what they actually said and they have no problems correcting their errors. Thus depending on the corpus alone could yield the incorrect grammar. Further, Chomsky argues that native speakers have can easily understand and/or say sentences they have never heard or said before. A grammar has to reflect this creative property of human language differentiating it from other systems of communication. Thus, Chomsky argues that a linguist should aim at describing the speaker's mental grammar by eliciting his intuitions. He makes a fundamental distinction between competence and performance. For Chomsky, competence is the linguistic knowledge a speaker has of his language while performance is what he actually says which a true reflection of his linguistic knowledge is not always. Chomsky [12] states clearly that linguistic theory must be concerned with characterizing native speakers' competence rather than performance.

Transformational Generative grammar [12] had interesting implications for computational linguistics and machine translation. First, because of the creativity of language, a grammar of a language has to make a distinction between an infinite set of sentences representing what has been said and could be said in that language and ungrammatical sentences. Because languages are learnable, their grammar must be finite, thus a grammar consists of a finite set of rules that generates an infinite set of sentences. Hence recursion has become an important property of phrase structure grammar. Second, the grammar that is elucidated by the linguist should mimic native speakers' intuitions, Since Chomsky points out those native speakers can recognize the different interpretations of an ambiguous sentence. Native speakers of Arabic can assign at least two different interoperations of the following sentence:

(1) قابلت مدير البنك الجديد

Speakers of Arabic would recognize that it can be translated as either 'I saw the new manager of the bank' or 'I saw the manager of the new bank'. An adequate grammar of Arabic must mimic native speakers' ability to recognize the ambiguity of such a sentence by assigning two different structural descriptions to the sentence. Because ambiguity is one of the most challenging aspects of NLP, computational linguists should understand the relevance of generative grammar to their work. For example, an Arabic to English machine translation system must assign different structures to the seemingly identical sentences in (3) and (4).

(2) قرأت كتابا لتشومسكي
I read a book by Chomsky.

(3) أعطيت كتابا لتشومسكي
I gave a book to Chomsky.

The problem here is known as the prepositional attachment problem which is to identify when a preposition should attach to the verb or to the noun phrase. Moreover in the case of Arabic the correct translation of the preposition in cases like this depends on the relationship of the preposition to the constituent it modifies. Figures (1) and (2) below show that the structure of (2) must be different from that of (3); thus capturing the difference in the semantic interpretation of the two.

One may substitute the constituent 'كتابا لتشومسكي' "a book by Chomsky" in (2) by one word such as " a story" as in (4) whereas doing so in (3) results in an ungrammatical sentence such as in (5).

(4) قرأت قصة
(5) * أعطيت كتابا قصة

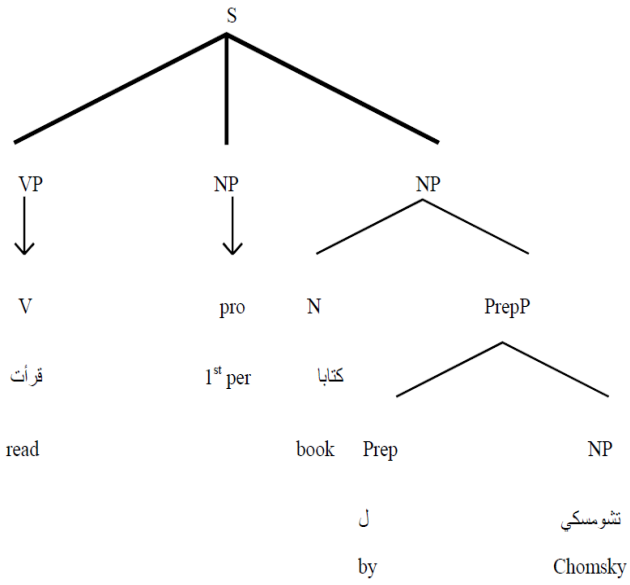

Fig. 1. Sentence (2) where the PrepP is modifying the noun

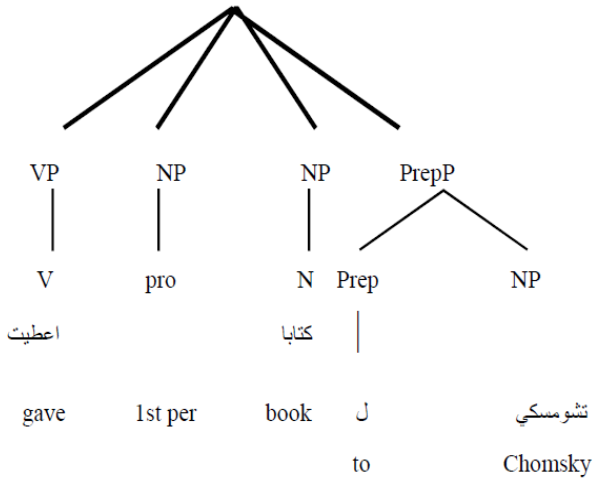

Fig. 2. Sentence (3) where the PrepP is modifying the verb

The possibility of substituting the phrase 'a book by Chomsky' by one word suggests that the phrase is one constituent as the tree in Figure (1) shows. In Figure (1) the PrepP is dominated by the NP which shows it is sub-part of the NP. In Figure (2) the PrepP is dominated by S which shows it is a constituent as the same level as the direct object of the sentence. Such analysis is crucial for the correct translation of the sentence as a whole and for the preposition in particular.

C. Progress in NLP Technology

As linguists became more concerned with producing formal descriptions of the languages they studied, it became apparent that the computer was an indispensable tool to their work. A formal description of a language must produce and analyze sentences that native speakers accept as paraphrases of each other or as being ambiguous [14]. The validity of the formal description rests on its ability to produce the same judgments a native speaker makes about his language. The computer was an excellent tool that forced linguists to be more explicit and precise in their formal description of natural languages. So more linguists became interested in using the computer to test the formal descriptions and the grammars they derived to analyze and generate natural language.

This created a further need to make programming more accessible to linguists. For example, linguists were trained to write phrase structure grammar (PSG) which is a form of Context Free Grammars (CFG). They strictly followed the Formal language theory in their PSG. The distinction between terminal and non-terminal symbols and was crucial in the development of chart parsing [14]. Chart parsing combines the best of top-down and bottom-up parsing. A chart parser takes as an input a string of words and runs several procedures to generate a chart representing the syntactic structure of the input sentence. Other formalisms were introduced that had a great impact on the progress in Computational Linguistics such as the development of Definite Clause Grammar (DCG) by Fernando Pereira [15] and Unification Grammar by Stuart Shieber [16].

The progress in linguistic formal descriptions of natural languages and the availability of advanced dedicated computational technology for natural language processing provided a stimulating environment which promoted what is known now as "deep parsing" and rule-based NLP.

D. The SYSTRAN Arabic to English Transfer Machine Translation System

SYSTRAN Inc. has been a pioneer in machine translation for over than thirty years focusing on developing machine translation systems for more than thirty languages using the transfer approach. While the direct method in MT involves two main stages, the transfer approach has three distinct stages: *analysis* of the source language; transfer of the structure of the source language to that of the target language; and the *generation* stage which produces the target language. SYSTRAN is also recognized for its use of extensive dictionaries that annotate lexical items with morphological, syntactic and semantic features. Since transfer machine translation systems are usually designed for specific language pairs, they can capitalize on the similarities between the source and target languages. They also use more sophisticated linguistic knowledge than that used in the direct method.

The development of the SYSTRAN Arabic to English machine translation system began in San Diego in June, 2002 initially, with a small grant from the US government. The author managed the project under the supervision of Jean Senellart, the Director of Research and Development at SYSTRAN.

Following, is a description of the development of this rule-based Arabic MT system beginning with the first phase; an Arabic gisting MT system.

1) The Gisting Phase

The funding agencies had an urgent need for a gisting system that uses unstructured, unvocalized, Arabic documents as input and generates a word-for-word English translation making it possible for someone with no knowledge of Arabic to intelligently guess the subject of the original Arabic document. This can be very valuable especially when the user is faced with enormous amounts of Arabic texts; and clearly, sorting potentially relevant from irrelevant documents saves both time and money. Further, it was required that the system be both fast and that coverage should not be less than 95 percent. To meet these stringent requirements, SYSTRAN developed a monolingual Arabic stem-based lexicon and a bilingual Arabic to English dictionary. To expedite the process, it was decided to use Arabic stems rather than roots, eliminating the step of generating stems from roots. Thus, each lemma is associated with a set of stems. For example, a lemma of an Arabic verb is associated with five stems: the perfect, imperfect, imperative, passive perfect and passive imperfect. Lexicographers were provided with the output of a guesser that generated all the required stems, with the additional requirement that the output of the guesser had to be validated and corrected. A morphological generator was also developed to generate all the inflected forms of the lemmas in the dictionary. With these components, coverage at the end of the first three months of the project was 80 percent. Continuous testing on the Aljazeera web site, Arabic newspapers and entering new words in the dictionary increased coverage to 96 percent by December, 2002.

2) Internal and External Arabic Morphology

The traditional Arab grammarians' account of Arabic morphology in terms of roots and patterns is very precise and explicit and since the 1980's, there has been extensive research on computational treatments of Arabic morphology [17] [18] [19] [20] [21]. Most work on Arabic morphology aims to identify and separate the prefixes and suffixes from the surface word and recover the root or the stem that may have undergone morphophonemic changes. But this is not a trivial problem for a computer program to solve. SYSTRAN made a fundamental distinction between two kinds of affixes that can be attached to Arabic stems and/or roots. The first type is the affix that has only a grammatical meaning such as subject-verb agreement markers, tense or mood markers. These affixes are not part of the SYSTRAN dictionary but are generated by SYSTRAN's Arabic morphological generator. This generator takes as input the list of stems and their part of speech tags from the dictionary and generates all the surface forms that each stem could assume. The result is a run time dictionary that has words as they actually occur in authentic Arabic unstructured texts. An example of these affixes is the regular masculine plural markers ون, ين and the regular feminine plural ات. At SYSTRAN's system, internal morphology is pivotal: it is where all different forms of one and only one stem are generated.

But with regard to external morphology, Arabic is an agglutinative language. Thus, affixes representing different parts of speech can be conjoined together with a stem or a root to form a token that has a syntactic structure. For example, the token بالمدينة `in the city' is a prepositional phrase that has a stem مدينة `a city' which is a noun, and two prefixes; the first is ب `in' which is a preposition and the second is ال `the' which is the definite article. Thus, external morphology describes the way the affixes that represent different parts of speech are attached to Arabic stems and the order of attachment is rule governed. SYSTRAN's Arabic external morphology defines the syntax governing the agglutination of Arabic complex words (see [22] for specific examples of the Arabic internal and external morphological rules).

3) Arabic Syntactic Analysis and Disambiguation

The goal of the second phase of the SYSTRAN Arabic to English MT system was to improve translation quality by introducing analysis, transfer and disambiguation rules. Several rules for recognizing noun phrases and their boundaries were introduced with transfer rules to transform the Arabic NP structure to the English structure. For example, a common Arabic noun phrase has the structure "Det Noun Det Adj" as in الرجل الطويل `the tall man' is transferred into the corresponding English structure Det ADJ Noun. Figure (3) below represents the syntactic structure of that phrase generated by SYSTRAN's analysis component while Figure (4) represents the output of the transfer component which maps the Arabic structure of noun phrase into it s corresponding English structure. Finally, the generation of the English phrase is accomplished by substituting the Arabic lexical items to their English equivalents as shown in Figure (5). Thus, the three trees represent the three stages: analysis, transfer and generation. Please note we have left out, for the simplicity of the analysis, the lexical and agreement features that are usually checked in the analysis.

Several analysis rules to recognize and correctly translate the Arabic genitive noun phrase known as the ``idaafa or noun construct" were introduced. Similarly, several rules for sentence structure were introduced to transfer the common Arabic VSO structure into the SVO English word order. The implementation of the analysis and transfer rules, though

limited, resulted in a marked improvement in translation quality.

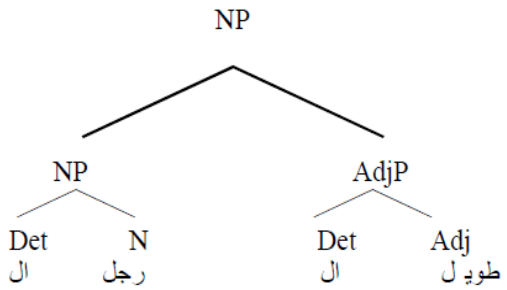

Fig. 3. The output of the analysis component

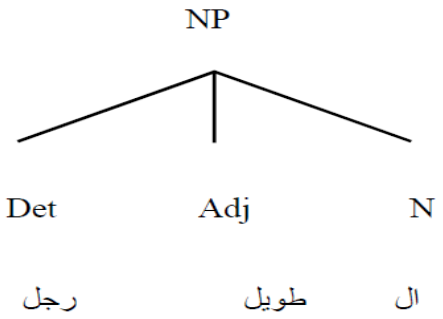

Fig. 4. The output of the transfer component

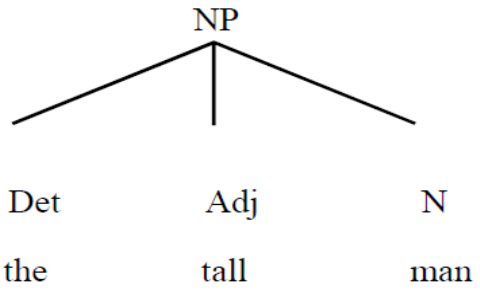

Fig. 5. The output of the generation component

Another improvement was achieved through homograph resolution and word sense disambiguation. In dealing with homograph resolution, we found that the most frequent homograph ambiguity was between nouns/adjective (almost ninety percent of the observed ambiguities). This high degree of noun-adjective homograph ambiguity arises from the nature of the structure of the Arabic language. In Arabic, adjectives and nouns inflect in the same way. Like Arabic nouns, an Arabic adjective inflects for gender, number, case, and definiteness. It is not surprising, then, that traditional Arabic grammarians subsume adjectives under nouns and consider that there are only three main parts of speech in Arabic: nouns, verbs and particles. SYSTRAN implemented contextual rules for homograph resolution. For example a noun/verb ambiguity is resolved as a noun if the ambiguous word or phrase is preceded by a preposition because Arabic does not allow prepositions to precede verbs. SYSTRAN also implemented contextual rules with look-ahead and look-back features for word sense disambiguation. For example, in the absence of diacritization the Arabic verb يزور could be translated as `visit' or `forge'. Our word sense disambiguation module would look ahead to see if it finds a noun with the feature `PLACE'. If so, the preferred translation would be `visit' rather than `forge' since in real life you do not "forge a place". It is more likely that you may visit a place. The word sense disambiguation module improved the quality of translation significantly.

E. Interlingua Machine Translation

The Interlingua approach to machine translation is based on the assumption that it is possible to convert the source language texts into a universal representation that is language independent. This universal representation can, in turn, be converted into the surface representation of the target language. While the transfer approach to machine translation is usually designed for a specific pair of languages, the Interlingua has the advantage of making the addition of a new language to the MT system less costly and much faster. Therefore, the Interlingua approach is more suited to multilingual environment. It was logical for the Eurotra machine translation project which was established and funded by the European Commission from the late 1970s until 1994 to adopt the Interlingua approach. According to the mandate of the Commission, all citizens of the European Union had the right to read and all the documents of the commission in their own language. With more countries joining the European Union (EU), this resulted in a combinatorial explosion in the number of language pairs involved and very quickly translation placed a heavy burden on the administrative budget of the EU. The Eurotra project aimed to solve this problem.

IV. Statistical Machine Translation

The Statistical Machine Translation approach is based on finding the most probable translation of a sentence using data gathered from an aligned bilingual corpus. Statistical machine translation has been gaining momentum in the last few years and there are several factors that make improving statistical MT systems faster and easier. First, the monolingual and bilingual data on the Web is growing providing enough data for language modeling and bilingual text alignment. Second, making MT system freely available on the Web (e.g. www.google.com) provides valuable crowdsourcing feedback to the systems. Third, academic research on statistical MT systems has grown which has already resulted in marked improvement. Fourth, current statistical MT systems do not suffer from fluency in the output which is still a problem for rule-based MT. Therefore more and more users prefer statistical MT systems over rule-based systems.

A. *The Language Weaver Arabic to English MT system*

Professors Kevin Knight and Daniel Marcu of the University of Southern California (USC) founded The Language Weaver Inc. in January 2002. The goal was to apply their pioneering research in statistical natural language processing to the commercial objective of producing useful automated machine translation systems. Fraser and Wong describe [23] one of the very first products that came from this remarkable transfer of academic research to industry. They present a detailed description of the creation of a complete statistical Arabic to English machine translation system. It is an excellent example of how rapidly and inexpensively a statistical MT system can be built when parallel corpora and training data are available. The technology behind statistical MT is presented and they provide examples of the various translation steps. The authors conclude the chapter with some suggestions for improving the Language Weavers' Arabic MT system. They cite the possibility of incorporating some statistically-based syntactic analysis, more sophisticated morphology, a special module for treating transliterated names of persons and companies, and delivering a ``learning" module that the customer could use after post-editing the translation output. Empowering users with features allowing them to modify the translation engine to better serve their specific domain and to correct some observed translation inaccuracies, is an excellent addition to the current systems.

B. *The AppTek Hybrid MT System*

Sawaf [24] describes a hybrid MT system for translating written and spoken MSA texts as well as Iraqi Arabic. Sawaf [24] reviews the two main approaches to machine translation: statistical and rule-based. After carefully evaluating the advantages and disadvantages of each approach, he presents the Apptek MT system (http://apptek.com/), an embodiment of the positive features in both approaches.

There are three innovations in the Apptek MT system. First, Sawaf found that the translation of entities such as person names and dates using a named entity recognition component improves the quality of the translation. Once a named entity is recognized, Apptek uses several approaches to translate such entities. For examples a token such as أمل when recognized as a person name, would not be translated into its linguistic meaning `hope'. Rather it would be transcribed into its phonetic representation in the target language /Amel/. Secondly, Apptek's system incorporates Arabic dialects. For example, the current system translates the Iraqi dialect into Modern Standard using a bilingual corpus, linguistic features of both MSA and the Iraqi dialect, and data training sets. Thirdly, it combines Arabic speech recognition output with machine learning. The speech recognition engine was trained using corpora in MSA and the Iraqi dialects.

V. CONCLUSION

We traced the beginnings of machine translation from the vision that Descartes had 400 years ago to the first realization of his dream in the twentieth century. Since then, machine translation technology has evolved through at least three generations starting from the Direct Method, followed by the Transfer Approach, which was succeeded by the Statistical MT approach. We briefly described three Arabic machine translation systems, which were developed following one of the three approaches. Thus, Arabic machine translation has been part and parcel of main stream machine translation and as such, it has undergone the same development paradigms as mainstream MT.

REFERENCES

[1] Zaennen, Annie. 2006. Mark-up Barking Up the Wrong Tree. Computational Linguistics 32(4):557-580.

[2] Reiter, Ehud. 2007. The Shrinking Horizon of Computational Linguistics. Computational Linguistics 33(2):283-287.

[3] Jones, Karen. 2007. Computational Linguistics: What about the Linguistics. Computational Linguistics 33(3):437 – 441.

[4] Benajiba, Yasin and Imed Zitouni. 2009. Morphology-Based Segmentation Combination for Arabic Mention Detection. ACM Transactions on Asian Language Processing 8(4) Article 16.

[5] Sawaf, Hassan. 2010. The AppTek Hybrid Machine Translation System. in Arabic Computational Linguistics, edited by Ali Farghaly, CSLI Publications:

[6] Fraser, Alexander and William Wong. 2010. The Language Weaver Arabic to English Machine Translation System, in Arabic Computational Linguistics, edited by Ali Farghaly, CSLI Publications: 251 – 282

[7] Hutchins, John. 2000. Early Years in Machine Translation, Ed. John Hutchins, John Benjamin, Amsterdam.

[8] Garvin, Paul. 1967. The Georgetown-IBM Experiment of 1954: An Evaluative Perspective. In Austin William M. (ed) *Papers in linguistics in honor of Dostert* (The Hague: Mouton, 1967), 46-56

[9] Hutchins, John. 1995. MACHINE TRANSLATION: A BRIEF HISTORY in *Concise history of the language sciences: from the Sumerians to the cognitivists*. Edited by E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press, 1995. Pages 431-445.

[10] ALPAC 1966 *Language and machines*: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee. National Academy of Sciences, Washington, DC.

[11] Chomsky, Noam, 1957. Syntactic Structures.Mouton, The Hague.

[12] Chomsky, Noam, 1965. Aspects of the Theory of Syntax., MIT Press, MIT Massachusetts.

[13] Bloomfield, Leonard, 1933. Language. New York: Holt, Reinhart and Winston.

[14] Kay, Martin. 1973. "The MIND System" in Rustin, Randall (ed.) Natural Language Processing, New York, the Algorithmics Press.

[15] Pereira, F. and D. Warren 1980. *Definite clause grammars for language analysis*.

[16] Shieber, Stuart. 1986. Introduction to Unification Based App. CSLI, Stanford, CA.

[17] Yehia, Hlal. 1985. Morphological Analysis of Arabic Speech. In Conference on Computer processing of the Arabic Language, Kuwait.

[18] Geith, Mervat. 1985. Morphology, Syntax and Semantics. In Conference on Computer Processing of the Arabic Language, Kuwait.

[19] Beessley, Kenneth. 2001. Finite State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In Proceedings of Workshop on Arabic NLP, ACL, Toulouse, France.

[20] Saudi, Abdelhadi, Antal van den Bosch and Gunter Neumann. 2007. Arabic Computational Morphology: Knowledge Base and Empirical Methods. New York, Springer.

[21] Attia, Mohamed, 2005. Developing a Robust Morphological transducer Using Finite State Technology. In 8th Annual CLUK Research Colloquium, Manchester, UK.

[22] Farghaly, Ali, 2003. Intuitive Coding of the Arabic Lexicon. In Proceedings of the MT Summit IX, Workshop on Machine Translation for Semitic Languages; Issues and Approaches, New Orleans.

[23] Fraser, Alexander and William Wong, 2010. The Language Weaver Arabic to English Statistical Machine Translation System, in Farghaly, Ali (Ed.) Arabic Computational Linguistics, CSLI, Stanfford.

[24] Sawaf, Hassan, 2010. The AppTek Hybrid Machine Translation System, in Farghaly, Ali (Ed.) Arabic Computational Linguistics, CSLI, Stanford, CA.

Ali Farghaly Professor in Computational Linguistics, Senior Member of Technical Staff, Text Group, Oracle USA, CA; Adjunct Professor of Arabic Linguistics, Monterey Institute of International Studies, Monterey, CA, USA.
E-mail: afarghal@miis.edu