החוג למדעי המחשב
Department of Computer Science
אוניברסיטת חיפה
University of Haifa
The Edmond Benjamin de Rothschild Institute for Interdisciplinary Computer Science

Israeli Seminar on Computational Linguistics

ISCOL'01

University of Haifa

15 February 2001

Israeli Seminar on Computational Linguistics ISCOL'01

Editor: ***Shuly Wintner***
Department of Computer Science
University of Haifa
31905 Haifa, Israel

Preface

Computational linguistics and natural language processing are active research fields in Israel today, as well as popular areas of activity in industry. The Israeli Seminar on Computational Linguistics is a venue for exchanging ideas, reporting on work in progress as well as established results, forming extramural cooperations and advancing the collaboration between academia and industry. This is a continuation of the tradition that started in 1995 with a meeting at the Technion and ended after four successful meetings in 1996.
This year we are happy to have two guest speakers:

Erhard Hinrichs
Seminar für Sprachwissenschaft
University of Tübingen, Germany

Sergei Nirenburg
Computing Research Laboratory
New Mexico State University

Abstracts were solicited in all areas of computational linguistics and natural language processing, as well as adjacent areas (formal and theoretical linguistics, psycholinguistics, information retrieval and information extraction, text mining, knowledge representation, speech processing, etc.), including work under way, provided that they represent recent and original work of general interest to our audience. Submissions of abstracts reporting on work pertaining to Hebrew and Arabic were especially encouraged. We were delighted to receive a great number of submissions, of which twelve were selected for presentation, in addition to the two invited keynote lectures. This booklet contains the abstracts of all the presentations.

The organization committee of this year's meeting consisted of Shuly Wintner, Department of Computer Science, University of Haifa, and Yoad Winter, Department of Computer Science, Technion. We gratefully acknowledge the financial support of The Edmond Benjamin de Rothschild Institute for Interdisciplinary Computer Science.

Shuly Wintner
University of Haifa

February 2001

Program

9:30-10:00	Registration and coffee
10:00-10:15	Opening
10:15-11:00	Invited talk: **Erhard Hinrichs**
	Robust Syntactic Annotation of Corpora and Memory-based Parsing
11:00-11:15	Break
11:15-12:55	Session I (4×25 minutes): **Applications**
	David Carmel, Doron Cohen, Miki Herscovici and Yoelle Maarek
	Palm Pirate - An Information Retrieval System for the Palm platform
	Lev Finkelstein, Envgeniy Gabrilovich, Zach Solan and Eytan Ruppin
	Placing Search in Context: The Concept Revisited
	Nissim Francez, Yaroslav Fyodorov and Yoad Winter
	An Order-Based Inference System for a Natural Language
	Svetlana Sheremetyeva and Sergei Nirenburg
	A Domain-Tuned Tool For Multilingual Document Management
12:55-14:00	Lunch break
14:00-15:40	Session II (4×25 minutes): **Hebrew and Arabic**
	Outi Bat-El
	On the Site of Vowel Deletion in Modern Hebrew Verbs and Nouns
	Gabi Danon
	The Hebrew Object Marker as a Type-shifting Operator
	Shlomo Izre'el
	Towards the Compilation of The Corpus of Spoken Israeli Hebrew
	Judith Rosenhouse
	Computerized Arabic Morphology and Syntax (for Teaching Purposes)
15:40-15:55	Break
15:55-17:35	Session III (4×25 minutes): **Statistical approaches**
	Meni Adler and Miki Tebeka
	Unsupervised Hebrew Part-of-speech Tagging
	Yuval Krymolowski
	Augmenting Evaluation Schemes Using the Distribution of Performance
	Yuval Krymolowski and Ido Dagan
	Compositional Memory-based Partial Parsing
	Zvika Marx and Ido Dagan
	Conceptual Mapping through Keyword Subset Coupling
17:35-17:50	Break
17:50-18:35	Invited talk: **Sergei Nirenburg**
	Ontological Semantics and Its Applications

Contents

Robust Syntactic Annotation of Corpora and Memory-based Parsing
Erhard Hinrichs 1

Palm Pirate - An Information Retrieval System for the Palm platform
David Carmel, Doron Cohen, Miki Herscovici and Yoelle Maarek 3

Placing Search in Context: The Concept Revisited
Lev Finkelstein, Evgeniy Gabrilovich, Zach Solan and Eytan Ruppin 5

An Order-Based Inference System for a Natural Language
Nissim Francez, Yaroslav Fyodorov and Yoad Winter 7

A Domain-Tuned Tool For Multilingual Document Management
Svetlana Sheremetyeva and Sergei Nirenburg 9

On the Site of Vowel Deletion in Modern Hebrew Verbs and Nouns: A Constraint-based Approach
Outi Bat-El 11

The Hebrew Object Marker as a Type-shifting Operator
Gabi Danon 13

Towards the Compilation of The Corpus of Spoken Israeli Hebrew
Shlomo Izre'el 15

Computerized Arabic Morphology and Syntax (for Teaching Purposes)
Judith Rosenhouse 17

Unsupervised Hebrew Part-of-speech Tagging
Meni Adler and Miki Tebeka 19

Augmenting Evaluation Schemes Using the Distribution of Performance
Yuval Krymolowski 21

Compositional Memory-based Partial Parsing
Yuval Krymolowski and Ido Dagan 23

Conceptual Mapping through Keyword Subset Coupling
Zvika Marx and Ido Dagan 25

Ontological Semantics and Its Applications
Sergei Nirenburg 27

Robust Syntactic Annotation of Corpora and Memory-based Parsing

Erhard Hinrichs (eh@sfs.nphil.uni-tuebingen.de)

Seminar für Sprachwissenschaft
Eberhard-Karls-University, Tübingen

This talk will provide an overview of current research at the SfS on the syntactic annotation of the VERBMOBIL corpus of spoken German and the German reference corpus (DEREKO) of written texts. Syntactic annotation for these corpora is performed automatically by a hybrid architecture that combines robust symbolic parsing with finite-state methods (”chunk parsing” in the sense Abney) with memory-based parsing (in the sense Daelemans).
The resulting robust annotations can be used by theoretical linguists, who are interested in large-scale, empirical data, and by computational linguists, who are in need of training material for a wide range of language technology applications.

Palm Pirate - An Information Retrieval System for the Palm platform

David Carmel, Doron Cohen, Miki Herscovici and Yoelle Maarek
({carmel,doronc,miki,yoelle}@il.ibm.com)

IBM Research Lab in Haifa

Personal Digital Assistants (PDAs) have been rising in popularity in the last few years. As their storage capabilities continue to improve, PDAs are turning into reference tools - containers for small (several megabytes) to medium-size (hundreds of megabytes) collections of short documents. Candidate user communities include medical students, physicians, lawyers and others. Adequate searching mechanisms for these collections thus become a necessity.

We introduce here Pirate Search for Palm (nickname Palm Pirate), a full text information retrieval system for Palm-like devices. Palm Pirate supports all of the state-of-the art features of modern search engines such as: stemming, scoring and ranking for great accuracy while achieving fast response time in spite of the platform limitations. With Palm Pirate, one can search not only local memos, but also larger textual collections. While the built-in PalmOS ”find” uses sequential string search and is therefore both inaccurate and too slow on collections of more than a few dozens documents, Palm Pirate indexes the document collection on the PC companion, and copies the generated index at synchronization time so as to guarantee fast constant response time and top accuracy whatever the size of the collection is.

From an implementation viewpoint, the main innovation lies in the indexing being conducted on the desktop companion rather than on the PDA, and replicate the inverted index on the PDA at synch time. From a technical viewpoint, the key achievements of Palm Pirate are its indexing and storage algorithms that allow for a small index size and very fast subsecond response time.

The small index size eases on the space requirements of the device: 10% of the original collection instead of the standard 40%. For example, the King James Bible data takes 3MB and the added index is of only 300KB. This is achieved by using minimal perfect hashing techniques.

Palm Pirate has been developed at the IBM Research Lab in Haifa and is being offered for free download at http://www.alphaworks.ibm.com. Ongoing work for future releases includes index compression to achieve even lower space overhead and Hebrew support so as to support search of classical Hebrew reference texts such as the ”tanakh”. The support for search in Hebrew will take advantage of previous work conducted by the group in the domain: namely Hemed (presented at NGITS 1999) an original statistical morphological disambiguator for Hebrew developed specifically for Hebrew search engines. We will discuss here the challenges posed by using a morphological analyzer rather than a simple stemmer (as done in English) in light of the limited storage capacities and our first attempts at solving this problem.

Placing Search in Context: The Concept Revisited

Lev Finkelstein, Evgeniy Gabrilovich, Zach Solan and Eytan Ruppin
({lev,gabr,zach,eytan}@zapper.com)

Zapper Technologies Inc.

Given the constantly increasing information overflow of the digital age, the importance of information retrieval has become critical. Web search is today one of the most challenging problems of the Internet, striving at providing users with search results most relevant to their information needs. Search engines have now entered their third generation, and current research efforts continue to be aimed at increasing coverage and relevance.

A large number of recently proposed search enhancement tools have utilized the notion of *context*, making it one of the most abused terms in the field, referring to a diverse range of ideas from domain-specific search engines to personalization. We present here a novel search approach that interprets context in its most natural setting, namely, a body of words surrounding a user-selected phrase. We postulate that a large fraction of searches originate while users are reading documents on their computers, and require further information about a particular word or phrase. Hence, the basic premise underlying our approach is that these searches should be processed *in the context of the information surrounding them*, allowing more accurate search results that better reflect the user's actual intensions. For example, a search for the word "Jaguar" should return car- related information if performed from a document on the motoring industry, and should return animal-related information if performed from an Internet website about endangered wildlife.

The significance of the new *context-based* approach lies in the greatly improved relevance of search results. Existing approaches either analyze the entire document the user is working on, or ask the user to supply a category restriction along with search keywords. As opposed to these, the proposed method analyzes the context in the immediate vicinity of the focus text, without running over the more distant (and less related) topics in the source document. The method also allows collecting contextual information *autonomously* without conducting an explicit dialog with the user.

Our system (named IntelliZap) is based on the client-server paradigm, where a client application running on user's computer captures the context around the text highlighted by the user. The server-based software uses a novel high-dimensional clustering algorithm to select the most important semantic themes, and then prepares a set of augmented queries based on them. Queries resulting from context analysis are dispatched to a number of general search engines, performing meta-searching. When the context can be reliably classified to a predefined set of domains (such as health, sport or finance), additional queries are dispatched to search engines specializing in this domain. A dedicated *reranking* module ultimately reorders the results received from all the engines, according to semantic proximity between their summaries and the original context.

Both the clustering and the reranking algorithms use a semantic network for measuring distances between pairs of words. To this end we developed a semantic metric that given a pair of words or phrases returns a (normalized) score reflecting the degree to which their meanings are related. Our semantic network merges statistical information on word cooccurrences in text corpora with linguistic information based on WordNet. The former component represents words as vectors in a multi-dimensional space, similarly to the canonical information retrieval approach. The latter component computes word similarity using the information content criterion estimated on the WordNet hypernymy hierarchy.

An Order-Based Inference System for a Natural Language

Nissim Francez, Yaroslav Fyodorov and Yoad Winter
({francez,yaroslav,winter}@cs.technion.ac.il)

Technion

In our work we develop a version of Natural Logic – an inference system for a natural language that works directly on natural language syntactic representations, with no intermediate translation to logical formulae.
Model-theoretic semantic theories of natural language assume that most linguistic expressions – or even all of them – represent objects in partially ordered domains so that meanings of expressions of the same category are naturally comparable. Formal semantics treats order relations between expressions of complex categories as compositionally derived from orders between expressions of simpler categories, according to the structural rules of a given grammar and certain semantic properties of words.
Following work by Sanchez, we develop a small fragment that computes semantic order relations between derivation trees in Categorial Grammar. The semantic information for the inferences is provided by special semantic markers on the syntactic categories that the Categorial Grammar derives for natural language expressions. Unlike previous works, the proposed system has the following new characteristics: (i) It uses orderings between derivation trees as purely syntactic units, derivable by a formal calculus, that we call the Order Calculus. In the Order Calculus each formula is a pair of derivation trees in Categorial Grammar. (ii) The system is extended for conjunctive phenomena like coordination and relative clauses. This allows a simple account of non-monotonic expressions that are reducible to conjunctions of monotonic ones. Such a treatment is possible using a treatment of coordination in natural logic that relies on the semantic fact that items like 'and' and 'or' are greatest lower bound/least upper bound operators respectively with respect to the order relations in the categories they apply to.
Another question about natural logic is whether variants of this system are decidable. We provide a partial answer to this question by the development of a preliminary proof search algorithm for the Order Calculus that is sound and is complete under certain semantic assumptions for a subset of the inference rules. The current proof search algorithm has an exponential (at least) complexity.
This work is currently under way and we are aiming towards finding an algorithm for the complete set of the inference rules (or proving its undecidability).

A Domain-Tuned Tool For Multilingual Document Management

Svetlana Sheremetyeva and Sergei Nirenburg ({lana,sergei}@crl.nmsu.edu)

Computing Research Laboratory
New Mexico State University

An approach to developing an interactive domain-tuned MT tool for multilingual document management is described. The tool for translating patent claims between Russian and English which is a part of a workstation for multilingual processing of patent texts is taken as an example.

General purpose MT systems have the advantage of being potentially reusable, this reusability is not however guaranteed. For example, no existing NLP system can process patent texts adequately. It is generally recognized, however, that an MT system providing adequate performance even for a single type of texts should be considered useful. Indeed, practically all MT systems for special domains are usually built tuned to the constraints of a sublanguage. Our approach conforms to the human-aided machine translation paradigm. In our model the initiative is dominantly if not exclusively, with the system.

The tool is meant for a SL speaker who does not know the TL. It consists of a) an analysis module which performs interactive syntactic analysis (decomposition of a complex nominal sentence into a set of simple structures) and fully automated morphological analysis of the word occurrences in these simple structures, b) an automated module for transferring the lexical and partially syntactic content of SL text into a similar content of the TL text, and c) a fully automated TL text generation module which relies on knowledge about the legal format TL patent claims.

An interactive analysis module guides the user through a sequence of SL analysis procedures, as a result if which the system produces a set of internal knowledge structures which serve as input to the TL text generation. Both analysis and generation rely heavily on the sublanguage of patent claims. The model has been developed for English and Russian as both SLs and TLs but is readily extensible to other languages.

On the Site of Vowel Deletion in Modern Hebrew Verbs and Nouns: A Constraint-based Approach

Outi Bat-El (obatel@post.tau.ac.il)

Tel-Aviv University

There are quite a few nouns and verbs in Modern Hebrew that look alike, to the extent of near minimal pairs like šafan 'rabbit ms.' and šamar 'to guard', as well as a few minimal pairs such as gamal 'camel' - gamal 'to reward'. Despite the surface identity, these nouns and verbs are phonologically distinct. When a vowel initial suffix is added to the stem, a noun deletes the first stem vowel (gamal - gmalim) while a verb deletes the second stem vowel (gamal - gamlu).

I will argue that this distinction is due to one constraint which holds for verbs but not for nouns (I ignore nouns that do not undergo vowel deletion). The argument is given within the framework of Optimality Theory (Prince and Smolensky 1993) which is based on constraint interaction. The constraints are universal, and hierarchically organized on a language-specific ground. A constraint can be violated when in competition with a higher ranked constraint (and thus of higher priority).

I will claim that vowel deletion is triggered by the constraint FOOT BINARITY (FTBIN), whose function in the language is reflected by a tendency towards a foot-size word (i.e. two syllables). Since the morphological operation (suffixation) must be surface true (due to the dominance of the morphological constraint which attaches the affix), there are two ways to satisfy FTBIN, deletion of the first or the second stem vowel. Deletion of the first vowel results in a syllable with a complex onset (CCV), while deletion of the second vowel results in a syllable with a coda (CVC). Both syllables are universally marked, violating constraints against complex onset (*CXONSET) and coda (*CODA) respectively.

I will propose that the relevant constraints are ranked as follows: FTBIN >> *CXONSET >> *CODA (where A>>B means that A has a priority over B). However, *CXONSET is a family of constraints consisting of *CXONSET[VERB] which is relevant to verbs only, and a general *CXONSET. The more specific constraint is ranked above *CODA, while the general one below *CODA, giving the ranking FTBIN >> *CXONSET[VERB] >> *CODA >> *CXONSET. When the input is either a verb or a noun, FTBIN rules out an output which has not undergone deletion, as it contains three syllables (*gamalim, *gamalu). When the input is a verb, *CXONSET[VERB] rules out the candidate with the complex onset (*gmalu); the winner is then the candidate with the coda (gamlu). When the input is a noun it goes freely through *CXONSET[VERB], allowing *CODA to rule out the candidate with the coda (*gamlim); the winner is then the candidate with the complex onset. Using this ranking as a base, I will also explain why some verbs (denominatives) violate *CXONSET[VERB] (e.g. tilgref), and why some nouns (actually adjectives) violate *CODA (e.g. tipeš - tipša 'stupid ms.-fm.'). Throughout the paper I will emphasize the superiority of the constraint-based approach adopted here over a rule-based approach which would have had to provide two different deletion rules, one for nouns and another for verbs. In particular, I will point out that the rule-based approach does not reflect the generalization that the two types of vowel deletion serve the same function, i.e. satisfying FTBIN. I will also consider an alternative approach whereby the site of deletion is encoded in the suffixes, and provide two reasons to reject such an approach. One is the dual function of the suffix -a, which can attach to both verbs and nouns. The other is that *CXONSET serves also in distinguishing between suffixed nouns and acronym words, where in the latter *CXONSET is never violated. I will suggest that the dichotomy of *CXONSET to various sub-constraints stems from the fact that Modern Hebrew phonology is a mixture of two phonologies, that of Tiberian Hebrew where *CXONSET is high-ranked, and that of Slavic languages where *CXONSET is low-ranked.

The Hebrew Object Marker as a Type-shifting Operator

Gabi Danon (danon@post.tau.ac.il)

Department of Linguistics, Tel-Aviv University

It is well known that the Hebrew object marker et usually precedes definite objects only. However, it has also often been observed (Glinert 1989, Ziv 1982 and others) that the distribution of et cannot be accounted for in purely semantic terms: There are semantically definite DPs, such as sefer ze ('this book'), which cannot follow et, as well as indefinites like axad ha-sfarim ('one of the books') which may follow et. For the vast majority of DPs in Hebrew, the presence or absence of et can be predicted from the syntactic structure of the object and from formal definiteness marking of its head.

This talk will focus, however, on a small number of cases where the use of et is optional. This includes certain quantified DPs, the question word ma ('what'), and conjunctions of two or more definite DPs, where et can optionally precede each conjunct (Winter 2000). In all these cases, the presence of et affects the interpretation of the sentence. However, I argue that the semantic contribution of et is not directly related to definiteness. In certain cases, such as when et precedes an object like axad ha-sfarim, the use of et produces a "specific" interpretation of the partitive (see En 1991 for similar facts in Turkish); and when et precedes the question word ma ('what'), an appropriate answer would only be a definite or a specific DP, even though it would be unnatural to claim that the wh- word itself is definite. But most strikingly, when the object is a conjunction of two definites, the repetition of et in front of every conjunct has nothing to do not only with definiteness but also with specificity: As discussed in Winter (2000), multiple occurrences of et give rise to a distributive reading, as opposed to a preference for a collective reading of the conjunction otherwise.

I will propose that all these semantic effects of et can be accounted for by the assumption that et is the overt realization of a standard lifting operator. Assuming, following Partee (1987), that there are DP denotations at each of the 3 semantic types e, $< e,t >$ and $<< e,t >,t >$, it will be shown that the restriction that et imposes on the type of the DP which follows it correctly limits the range of available interpretations. Thus I argue that an analysis that makes use of all three semantic types for DPs, and which allows indefinites to have a type e interpretation, as in Reinhart (1997), can more easily account for the semantic effects of et than theories which make use only of a subset of these types, such as Barwise & Cooper (1981) and Winter (2000).

In the proposed analysis, definiteness and specificity are merely descriptive terms for two classes of DPs which happen to have a denotation of type e. Thus, the traditional notions of definiteness and specificity turn out to be only indirectly related to the observed facts, whose actual primary source is semantic type.

References

Barwise, Jon and Cooper, Robin (1981). Generalized Quantifiers and Natural Language. Linguistics and Philosophy 4, 159-219.

En, Mrvet (1991). The Semantics of Specificity. Linguistic Inquiry 22, 1-25.

Glinert, Lewis (1989). The Grammar of Modern Hebrew. Cambridge: Cambridge University Press.

Partee, Barbara (1987). Noun Phrase Interpretation and Type Shifting Principles. In Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers, Jeroen Groenendijk, Dick de Jong and Martin Stokhof (eds.). Dordrecht: Foris.

Reinhart, Tanya (1997). Quantifier Scope- How Labor is Divided between QR and Choice Functions. Linguistics and Philosophy 20, 335-397.

Winter, Yoad (2000). DP Structure and Flexible Semantics. Manuscript. Technion.
Ziv, Yael (1982). On so-Called ’Existentials’: A Typological Problem. Lingua 56, 261-281.

Towards the Compilation of The Corpus of Spoken Israeli Hebrew

Shlomo Izre'el (izreel@post.tau.ac.il)

Tel Aviv University

During the last two decades, much attention has been given to the methodology of corpus linguistics. In the main, using this methodology, linguistic description and theorizing is based upon statistical performance measures and observation of language use in real life. The first task is, obviously, the compilation of a body of texts, a corpus. Recent developments in computer sciences and the enormous change in computer storage capacity have greatly enhanced corpus studies throughout the world. While corpora have been compiled and continue to be compiled for many languages all over the world, there is still no available comprehensive corpus for modern Hebrew. Moreover, research on modern Hebrew, and especially on its spoken varieties, suffers greatly from the lack of descriptive studies, which is, among other issues, the result of a shortage of data.

A corpus – as we see it – is a preliminary desideratum for much larger projects that cannot otherwise be achieved, be it a grammar of modern Hebrew, a comprehensive dictionary, or any other theoretical or applied inquiry. The research potential a corpus represents is extremely large, and includes linguistic, cultural, sociological, and technological aspects.

With this in mind, we have decided to start procedures towards the compilation of The Coprus of Spoken Israeli Hebrew (CoSIH). The goals of the projects are:

1. To create a corpus of spoken Israeli Hebrew in order to facilitate research in a range of disciplines concerned with the Hebrew language and with the general methodology of Corpus Linguistics.

2. To disseminate this corpus publicly in multimedia format and in print. The multimedia format will be disseminated via electronic means including CD-ROM, DVD-ROM and the World Wide Web, and will present the recorded sound simultaneously with its transcriptions and other extensions, all linked together by software.

The design of CoSIH is set up as to include two complementary corpora: a main corpus and a supplementary corpus. The main corpus will form the bulk of CoSIH and will comprise about 90% of the entire collection. This will be a representative corpus, aiming at a representative sample of all varieties of Hebrew as it is spoken in Israel today, and will include representations of both demographic and contextual varieties. The supplementary corpus will include two distinct subcorpora: One will be based very much like the main corpus on demographic criteria, yet it will be compiled using non-proportional sampling. A second supplementary subcorpus will be compiled basically according to contextual criteria, with some attention paid to demographic features. Each of the distinct supplementary subcorpora will include about 5% of the entire corpus. In my lecture I will describe the preparations for the compilation of CoSIH, alongside the methodology of corpus linguistics, possible applications of the corpus, and especially its design, a pioneering effort to have a representative corpus which will integrate both demographic and contextual varieties of the spoken language.

Computerized Arabic Morphology and Syntax (for Teaching Purposes)

Judith Rosenhouse (gsrjudy@techunix.technion.ac.il)

Department of Humanities and Arts
Technion

Computational linguistics deals not only with theoretical questions but also with practical or applicative ones. The study of computational (or computerized) Arabic has been developing mainly for the last fifteen years at increasing pace. The main problems and goals referred to script (font) solutions, morphology and syntax, more or less as in studies of other languages. Language-specific problems include directionality, the richly inflected morphology and syntax-dependent word forms (case endings). In Arabic speaking countries this area has developed more slowly than in the West, where large diaspora-communities of readers of Arabic have been more exposed to the computers technology. The rapid parallel development of the Internet and other related advances in computer studies in the nineties have helped also Arabic in solving mainly problems of fonts as well as other linguistic problems.

This talk focuses on Arabic morphology and syntax from the point of view of their study in Israel. This aspect is important here, since Arabic is the 2nd official language of Israel and it is taught both as a mother tongue to its Arab native speakers, and as an obligatory foreign language to the Hebrew-speaking students, at least in the middle school. For this latter student body the Center of Educational Technology (Shif'at) has developed a system of programs (software and fonts) for 7th to 12th grade students, i.e., for all the middle school years and up to the end of high school.

First, a description of morphological and syntactic components of Arabic, through "computational eyes," will be presented. Examples will refer to points such as the following: (a) Morphology: root + pattern combinations for the noun and verb systems; noun patterns (including singular/dual/plural masculine/feminine sub-groups); verb patterns (measures, persons, tenses); (b) Syntax: concord (nouns + adjectives, nouns + verbs); word order (default; marked); negation; interrogation; active/passive transformations; and case endings. Then the CET-Shif'at system is briefly described, and some ideas about basic problems that require solutions and desirable future developments of programs are suggested.

Unsupervised Hebrew Part-of-speech Tagging

Meni Adler and Miki Tebeka ({adlerm,tebeka}@cs.bgu.ac.il)

Ben Gurion University

Part-of-speech tagging is the process of assigning grammatical categories to individual words in a corpus. The task of labeling each word with its appropriate part of speech is a well known problem in natural language processing. Several probabilistic approaches have been researched and applied in order to solve this problem, such as: hidden Markov model (Merialdo 1994), transformation-based learning (Brill 1995), decision trees (Schmid 1994), neural networks (Benello 1989), memory-based learning (Daelemans 1996) and maximum entropy models (Ratnaparkhi 1996). A supervised tagger is an implementation of a tagging model that uses an analyzed corpus as a learning basis. Designing an unsupervised tagger is more challenging, and is necessary where such a corpus does not exist (as is the case for Hebrew). Our work focuses on the unsupervised Hebrew tagging problem exploring two of the above approaches, hidden Markov model (HMM) and transformation-based learning.

Hidden Markov model

The tagging problem can be defined by HMM:
HMM is defined by (S,K, u) where:

S = s1,,sN a set of states.
K = k1, kM an output alphabet
u = (A,B,D) a probabilistic model where
D = i, i in S, initial state probabilities.
A = aij, i,j in S, state transition probabilities.
B = bijk, i,j in S, k in K, symbol emission probabilities.

In this model, the states correspond to tags and symbols emissions correspond to the words of the text. In the case of an unsupervised tagger, the probabilistic model is not known, and should be learned from an untagged corpus. The learning process is based on the Baum-Welch algorithm, which is a special case of the Expectation Maximization algorithm. The behavior of the algorithm under the English tagging problem was examined by Merialdo (1994) and Elworthy (1994). Merialdo compared the accuracy of an unsupervised tagger with a supervised one. Elworthy checked the effect of the initial conditions over the lexicon and the transitions. He also pointed the relation between the number of efficient iterations and the pattern of the re-estimation. In our work we examine the behavior of the Baum-Welch algorithm over Hebrew text, taking into account particularly the information that can be derived from the morphological analysis of the words.

Transformation based learning

In supervised transformational based tagging, the text is first annotated with the initial state anotator. Then, at each iteration, the result of the tagging is compared to the TRUTH. Transformations are learned that can be applied to the output of the initial state annotator to make it better resemble the TRUTH. The transformation are called patches and we use patch templates to create specific patches from them.
In the unsupervised version, the tagger in initial state tags every word with each of the possible tags given by a morphologic unit. Transformations are used to reduce the uncertainty as to the correct tag of a word in a given context. These contexts are specified by path templates. The tagger acquires several patches in the

form 'Change tag of a word from X to Y' where X is a set of tags and Y is a single tag. The scoring criterion is as follows:

1. freq(X) is the number of occurrences of words unambiguously tagged with X in the corpus.

2. incontext(Z,C) is the number of times a word unambiguously tagged with Z occurs in context C in the training corpus.

Let R = argmax[on Z] freq(Y) / freq(Z) * incontext(Z,C) then the score of 'Change tag of a word from X to Y' is: incontext(Y,C) - freq(Y) / freq(R) * incontext(R,C).
In each learning iteration the learner searches for the transformation which maximizes this functions and add this transformation to a path list.
After the learning stage tagging is done in the following manner:
1. Tag sentence with initial tagger
2. Apply each patch in turn.
In our work, we focus on particular parameters that are specific to the Hebrew language:

- The entropy of Hebrew, in term of tagging accuracy over a 'clean' unsupervised model, compared with English entropy.

- Hebrew tag set design: How should morphologic properties be used? What's the weight of the tag set design in the quality of the learning process?

- Initial condition: What are the main initial conditions that influence tagging accuracy?

- Integration of minimal hand tagging within the Baum-Welch algorithm.

- Unknown words: What heuristics can we use to tag unknown words?

[1] Elworthy David, 1994. Does Baum-Welch re-estimation help taggers? In ANLP 4, pp. 53-58.
[2] Merialdo, Bernard. 1994. Tagging English text with probabilistic model. Computational Linguistics 20:155-171.
[3] Benello Julian, Andrew W. Mackie, and James A. Anderson, 1989. Syntactic category disambiguation with neural networks. Computer Speech and Language 3:203-217.
[4] Daelemas, Walter, Jakub Zarvel, Peter Berck, and Steven Gillis, 1996. MBT: A memory-based part of speech tagger generator. In WVLC 4 pp. 14-27.
[5] Ratnaparkhi Adwait, 1996. A maximum entropy model for part-of-speech tagging. In EMNLP 1, pp.133-142.
[6] Schmid Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language processing, pp. 44-49, Manchester, England.
[7] L.E. Baum, T. Petrie, G. Soules and N. Weiss, 1970, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Annals of Mathematical Statistics, 41:164-171.
[8] Baum L., 1972, An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. Inequalities 3:1-8.
[9] Brill Eric, 1975, Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging, Computational Linguistics 21:543-565.
[10] Brill Eric, 1997, Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging To appear in Natural Language Processing Using Very LargeCorpora. Kluwer Academic Press.

Augmenting Evaluation Schemes Using the Distribution of Performance

Yuval Krymolowski (yuvalk@cs.biu.ac.il)

Bar-Ilan University

The performance of a statistical NLP system is commonly reported using a figure based on a single split between training and test data. Results obtained using a single split are subject to sampling noise. In this position paper we argue in favour of reporting a distribution of performance figures, rather than a single number. Such an evaluation scheme can be used for statistically quantifying statements concerning differences across parameter settings, systems, and corpora.
Using the bootstrap method, and a memory-based shallow-parsing algorithm, we show a marked difference between a test corpus which is close to the train corpus (WSJ), and a test corpus which comes from a different genre (ATIS).

Compositional Memory-based Partial Parsing

Yuval Krymolowski and Ido Dagan (yuvalk@cs.biu.ac.il)

Bar-Ilan University

We present a memory-based learning method that recognizes shallow patterns, including compositional ones (NP, VP), in a new text based on a bracketed training corpus. The examples are stored as-is, in efficient data structures. Generalization is performed at recognition time by comparing subsequences of the new text to positive and negative evidence in the corpus. The method is oriented for learning to parse any selected subset of target syntactic structures. It is local, yet can handle also compositional structures. Parts of speech as well as embedded instances are being used simultaneously. The output is a partial parse in which instances of the target structures are marked. Experimental results are presented for recognizing noun and verb phrases, as well as subject-verb and verb-object relations. We discuss analogies with DOP.

In essence, the method relies on storing sub-sequcences of parts-of-speech and embedded instances in a memory. In a new sentence, each word-range is tested as a pattern instance. In testing, the algorithm tries to reconstruct the POS sequence (incl. instances already found) from the tiles stored in memory. A constraint-propogation algorithm is invoked for choosing among conflicting alternatives at the *instance* level, no attempt is made to construct a parse for the whole sentence.

The embedding depth, and number of embedded instances in use, can be tuned - thereby simulating, e.g., to what extent is it necessary to go deep in hierarchy in order to learn a structure.

We are now working on producing a dependency parse, and might present preliminary results.

Conceptual Mapping through Keyword Subset Coupling

Zvika Marx and Ido Dagan (marxzv@cs.biu.ac.il)

Mathematics and Computer Science department
Bar-Ilan University

In this work, we introduce an extension to the well-studied task of measuring overall similarity of texts. We point out specific aspects of similarity, unique to comparison between distinct content worlds. For example, we would expect that in collections of news articles regarding various kinds of *conflicts* the notion of, say, *negotiating body* would be present. However, the vocabulary referring to this notion might vary: if the issue were an international conflict, the terms in use would be DIPLOMAT, DELEGATION and so on. On the other hand, for other types of conflicts the corresponding keywords could be JUDGE, LAWYER, COURT etc.

We have developed a new unsupervised learning framework, named *subset coupling*, that could be applied to domains in which context-dependent correspondences exist. Here we demonstrate its capabilities in matching context-dependently-related keyword sets extracted from distinct corpora. Following the natural identification of word clusters with concepts or conceptual categories, the obtained list represents mapping between conceptual entities presented by the two corpora.

Our work is inspired by classical cognitive theory of analogy (The *Structure Mapping theory*, Gentner, 1983). Criticism regarding the incompetence of the this approach in coping with real-world data (Hofstadter et al., 1995), is addressed by considering features of similarity that are salient in the context of mapping between the particular corpora.

We introduce an original algorithm for subset coupling, following a theoretical framework for cost-based pairwise clustering by Puzicha, Hoffman & Buhmann (2000). In order to capture the context of a particular comparison between data sets, our algorithm refers only to *between data-set similarities*, i.e. similarities between the keywords of one corpus to the keywords of the other one. This is a major difference from the conventional clustering scheme, which is devised for partitioning of one data set at a time. Cost-function optimization is achieved by standard relaxation method, namely the *Gibbs-Sampler* algorithm used also by Puzicha et al.

The keyword similarity values have been calculated according to a formula by Dagan, Marcus & Markovitch (1995). This measure gives significant weight to features that are observed with both terms for which the similarity value is calculated. In our case, these features are words used in both corpora (less a limited list of *stop words*). The conceptual mapping is thus guided by information regarding features and relations that are shared by the two particular systems under comparison.

As an illustrative case study, the distinct corpora, among which we have looked for correspondences, were focused on distinct *religions.* Keywords were (semi-automatically) extracted for each religion from the corresponding corpus (that is a set of introductory documents downloaded from the Internet). In the case of comparing Buddhism and Islam, our results include, among other items, correspondences between scripture-related keyword lists (PALI, SANSKRIT, ... – ARABIC, HADITH ...) as well as lists of terms referring to afterlife and reward (PAIN, REBORN, ... – JUDGMENT, PARADISE, ...).

We evaluate the results against data obtained from experts in the field of comparative religion studies, using conventional evaluation measures: accuracy and recall-precision curves.

References

Dagan, I., Marcus S. and Markovitch S. (1995) Contextual word similarity and estimation from sparse data. Computer Speech and language, 9/2, pp. 123-152.

Gentner, D. (1983) Structure-Mapping: A Theoretical Framework for Analogy. Cognitive Science, 7/2, pp. 155-170.

Hofstadter D. R. and the Fluid Analogies Research Group (1995) Fluid Concepts and Creative Analogies. Basic Books, New-York.

Puzicha J., Hofmann T. and Buhmann J. (2000) A theory of proximity based clustering: structure detection by optimization. Pattern Recognition 33/4 pp. 617-634.

Ontological Semantics and Its Applications

Sergei Nirenburg (sergei@crl.nmsu.edu)

New Mexico State University

I will introduce an application-oriented semantic theory that supports analysis and synthesis of texts through extracting and representing text meaning. The approach uses a constructed world model, or ontology, as its core knowledge component. Other knowledge components include lexica and onomastica for the natural languages used in an application as well as a fact database that stores instances of ontological concepts. Ontological semantics also introduces a detailed metalanguage known as the TMR, or ”text meaning representation”, language.

Ontological semantics has been used at NMSU CRL to support applications in machine translation, information extraction, question answering and task-oriented, collaborative human-computer environments.