Word Prediction in Hebrew – Preliminary and Surprising Results

Abstract

As part of an effort to develop NLP-based tools for Hebrew AAC users, we investigate the task of word prediction. Previous work on word prediction shows that statistical methods are not sufficiently precise for languages with highly inflected morphology, and that syntactic processing is required. Following this assumption, we have tested Hebrew natural language processing tools on the word prediction task. We found that while training a language model on a very large corpus (27M), we achieve high results on various genres including personal writing in blogs and in open forums in the Internet. Contrary to what we expected, using morpho-syntactic information such as part of speech tags decreases prediction results.

Background on Word Prediction

Word prediction aims at easing word insertion in textual software by guessing the next word, or by giving the user a list of possible options for the next word. A similar process happens naturally in human conversation between an AAC-user and a speaking partner. In such a situation, a speaking partner is most likely to predict the word that is to be said by using her knowledge about language and the context of the conversation [3]. The main purpose of word prediction is to speed up typing, but it can also help dyslexic people in reducing writing errors. This field of research has seen a surge of interest with the development of mobile phones (with their limited keyboard) and of hand-held devices.

The Typical Process of Word Prediction

Word Prediction is an active research field, the most recent survey on the subject appears in [3]. The main issues in the development of word prediction systems include the *prediction methods* and the *user interface issue,* which is out the scope of this work. Prediction methods include decisions on prediction units (characters, words), information sources and structure (both lexical and statistical), levels of linguistic processing, size and types of corpora and learning methods.

The process of prediction itself is based on:

1. Statistical information – mainly based on word frequencies, *i.e.,* taking into account probabilities of isolated words, or by more complex *Language Models* (LM) such as Markov models. Although unigrams are frequently used, trigrams yield better results

(but are harder to maintain). Word frequencies can be acquired from the user herself (using online learning methods), can be acquired from a corpus (a general or a domain-specific/user model one) or through a combination of both.

2. Syntactic knowledge - considering part of speech tags, and phrase structures. Syntactic knowledge can be statistical in nature or can be based on hand-coded rules (see [3] for a detailed description).

3. Semantic knowledge can be used by assigning categories to words and finding a set of rules which constrain the possible candidates for the next word. This method is not widely used in word prediction, mostly because it requires complex hand coding or may be time consuming and inefficient for real-time requirements. Recent work ([5], [10]) for instance) refines the prediction process by using semantic clues of related words that co-occur in texts.

All prediction methods require lexical data. Such data can be acquired from corpora along with word frequencies and lexical databases (which may be incorporated into the system). A word-prediction lexicon usually includes words frequencies. It may also include part of speech and semantic data. Lexicons can be adaptable, *e.g.*, updated with the user's vocabulary, and should be organized in an efficient way (linear vs. tree structure, with the trade-off of insertion cost).

Syntactic approaches require a set of linguistic tools such as POS taggers and lemmatizers, which are not available in all languages. Statistical methods are based on learning parameters from large corpora. This is problematic when the language that is written with the aid of the word prediction system is of a different style than the training data (which is, in most cases, obtained from newspapers).

Since the personal language that is used may be very different than the one that was used for modeling, systems must have a good strategy for handling unseen words or sequences of words (back-off models).

Some word predictors build their language model on-line and are updated as the user enters more text. This strategy is an effective way to balance the mismatch with "off the shelf" language models, but it suffers from the limited amount of data available to construct the individual language model.

Several heuristics are claimed to reduce the number of keystrokes significantly:

1. Recency promotion either by increasing statistical parameters of recently seen words, or by managing a file of the words used most recently.
2. The trigger and target method, where certain words can be used as a trigger to the possible presence of another word within some distance.
3. Capitalization of proper nouns and at the beginning of sentences (which is irrelevant for Hebrew).
4. Inflecting words where needed (based on syntactic knowledge).
5. Writing compounds (in languages with rich compounding like German or Dutch).
6. Distinguishing fringe and core words in the prediction process [9].

The drawbacks of the word prediction method lie mostly in the need to take an overt action to verify the system selection. Typing is, therefore, not a fluent task, which leads to additional cognitive load [6]. Still, [8] shows that word prediction can indeed increase AAC communication rate.

Evaluation of word prediction systems considers the keystroke savings, time savings, and cognitive overload (length of choice list vs. accuracy). A predictor is considered to be adequate if its hit ratio is high as the required number of selections decreases [3].

Word Prediction in Hebrew

Modern Hebrew is characterized by rich morphology, with a high level of ambiguity, in which several words combine into a single token in both agglutinative and fusional ways. Hebrew lexemes are derived according to various processes, such as composition (*e.g.,* of root and pattern), linear derivation (affixation and compounding), and null derivation [1]. Hebrew lexemes are inflected by gender, number, person, tense and construct state. These properties are applied to most content-carrying words. The definite article of Hebrew is attached pronominally to nominals: common nouns, adjectives, demonstratives, and numbers. A word in Hebrew can be attached, according to its lexical category, to a sequence of formative letters - מ,ש,ו,כ,ל,ב - indicating functions, such as coordination, preposition and subordination. In addition, nouns, adjectives, verbs, prepositions, and adverbs can be agglutinated with a pronominal pronoun suffix, indicating functions, such as possessive, accusative and nominative (with person/gender/number inflections) [1].

Previous work states that for languages with a rich inflectional morphology, a statistical method based on frequencies only is not sufficient, and a wide variety of syntactic knowledge is required [2] [3]. This assumption implies a mixed approach, which involves language models with part of speech information, as implemented in various systems (see references in [2]). For instance, predicting the first word of the sentence using a statistical model, then, parsing the partial sentence in order to predict the next words. Another possible method is by using two steps in prediction: one of a root and then of its possible inflections [3].

In this paper, we empirically evaluate the hypothesis that additional morpho-syntactic knowledge is necessary to obtain high-precision word prediction in Hebrew.

Experiment and Results

We calculate the keystroke savings as:

1-(# of actual keystrokes/#of expected keystrokes)

We test prediction on four sizes of selection menus: 1, 5, 7 and 9 (each considered as one additional keystroke), and we trained language models on unigrams, bigrams and trigrams.

In [4] it is shown that, the larger the corpus that the language model is trained on, the better predictions are achieved. Similar results were recently published by [7], who further show that a language model trained on a large corpus is more beneficial than a language model trained on small corpora of the same domain as the tested text.

Our results show a similar direction, however, we have trained the language model on various training sets of 1M, 10M and 27M words, which were found to achieve good results (Table 1). However, when we tested prediction performance using morpho-syntactic and syntagmatic information, the performance decreased (see Table 2). In this method, we re-rank the word candidates suggested by the language model, by calculating the probability of the sentence for each one of them according to the morpho-syntactic model: $P(w_n|w_1,\dots,w_{n-1}) = \lambda_1 P(w_{n-i},\dots,w_n|LM) + \lambda_2 P(w_1,\dots,w_n|\mu)$, where i is the order of the language model LM, and μ is the morpho-syntactic HMM model. For these experiments, we use $\lambda_1 = 0.8$ and $\lambda_2 = 0.2$.

NGram	Win	Test Set									
		News [90K words]			Blogs [50K words]			Med Forum [422K words]			Med Forum [53K words]
		Train Set Size			Train Set Size			Train Set Size			Train Set Size
		1M	10M	27M	1M	10M	27M	1M	10M	27M	364K same domain
3	1	0.71	0.63	0.54	0.78	0.75	0.74	0.77	0.73	0.72	0.62
	5	0.49	0.42	0.39	0.58	0.53	0.52	0.55	0.50	0.49	0.41
	7	0.46	0.39	0.34	0.54	0.49	0.48	0.51	0.46	0.45	0.46
	9	0.44	0.37	0.29	0.52	0.47	0.46	0.49	0.44	0.43	0.36
2	1	0.75	0.66	0.63	0.78	0.75	0.75	0.77	0.74	0.73	0.62
	5	0.50	0.44	0.41	0.58	0.54	0.53	0.55	0.50	0.49	0.41
	7	0.46	0.40	0.37	0.54	0.50	0.48	0.51	0.46	0.45	0.38
	9	0.44	0.38	0.36	0.52	0.47	0.46	0.49	0.44	0.43	0.36
1	1	0.80	0.79	0.79	0.85	0.84	0.83	0.84	0.83	0.82	0.73
	5	0.57	0.54	0.54	0.64	0.61	0.61	0.61	0.58	0.58	0.49
	7	0.53	0.50	0.50	0.60	0.58	0.57	0.58	0.55	0.54	0.38
	9	0.51	0.48	0.47	0.58	0.55	0.54	0.56	0.53	0.52	0.44

``````````````````````````````````**Table 1 Prediction results using n-grams only.**

| NGram | Win | Test Set | | | |
|---|---|---|---|---|---|
| | | News<br>[90K words] | Blogs<br>[50K words] | Med Forum<br>[422K words] | Med Forum<br>[53K words] |
| | | Train Set Size | Train Set Size | Train Set Size | Train Set Size |
| | | 1M | 1M | 1M | 364K<br>same domain |
| 3 | 1 | 0.75 | 0.78 | 0.86 | 0.78 |
| | 5 | 0.56 | 0.62 | 0.68 | 0.64 |
| | 7 | 0.53 | 0.61 | 0.67 | 0.61 |
| | 9 | 0.51 | 0.58 | 0.63 | 0.60 |
| 2 | 1 | 0.76 | 0.79 | 0.86 | 0.79 |
| | 5 | 0.57 | 0.62 | 0.69 | 0.64 |
| | 7 | 0.53 | 0.61 | 0.67 | 0.61 |
| | 9 | 0.51 | 0.59 | 0.64 | 0.60 |
| 1 | 1 | 0.82 | 0.83 | 0.91 | 0.83 |
| | 5 | 0.62 | 0.66 | 0.72 | 0.66 |
| | 7 | 0.60 | 0.66 | 0.70 | 0.63 |
| | 9 | 0.58 | 0.64 | 0.69 | 0.63 |

**Table 2 Prediction results with morpho-syntactic knolwedge.**

## Conclusions

This paper presents preliminary results of experiments in word prediction for Hebrew. The results are surprising: syntactic knowledge does not improve keystroke savings and even decreases them, as opposed to what was originally hypothesized.

As a starting point, statistical data on a language with rich morphology yields good results in comparison to other languages, saving up to 29% with nine word proposals and 34% for seven proposals, 54% for a single proposal.
``````````````````````````````````

In ongoing work, we address the issue of the influence of syntactic information on prediction results. We will also evaluate the model on a spoken Hebrew corpus, and address issues of error handling and prediction of named entities.

References:

[1] M. Adler (2007), *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. PhD thesis, Ben-Gurion University, Beer Sheva, Israel.

[2] P. Boissiere (2003). An overview of existing writing assistance systems. In Proceedings of the French-Spanish Workshop on Assistive Technology.

[3] N. Garay-Vitoria and J. Abascal (2005) Text prediction systems: a survey. Universal Access in the Information Society, 4:3, pp. 188-203.

[4] G.W. Lesher, B. J. Moulton, and D. J. Higginbotham (1999)Effects of ngrams order and training text size on word prediction. In RESNA 1999.

[5] J. Li and G. Hirst (2005), Semantic knowledge in word completion. In Proceedings of ASSETS'05, Baltimore, Maryland, October, 2005.

[6] S.M. Shieber and E. Baker (2003). Abbreviated text input. In IUI'03, Miami, Florida, USA.

[7] K. Trnka and K.F. McCoy (2007), Corpus Studies in Word Prediction. In Proceedings of ASSETS'07, Tempe, Arizona, October 2007.

[8] K. Trnka, D. Yarrington, J. McCaw, C. Pennington and K.F. Mccoy (2007), The Effects of Word Prediction on Communication Rate for AAC. In Proceedings of NAACL HLT 2007, Companion Volume, pp173-176, Rochester NY, Aril 2007.

[9] K.Trnka, D. Yarrington, K. McCoy, and C. Pennington (2006). Topic modeling in fringe word prediction for AAC. In *Proceedings of the 11th international Conference on intelligent User interfaces*, pp. 276-278 Sydney, Australia, January 2006.

[10] T. Wandmacher and J.Y. Antoine (2007), Methods to integrate a language model with semantic information for a word prediction component. In Proceedings of the EMNLP/CNLL, pp. 506-513, Prague June 2007.