Hebrew Syllabification

Project Summary

The project frame is: predicting the vowels of modern Hebrew text. The project optimizes a cascade of models from classical generative models up to a Recurrent Neural Network, all in the sequence labeling paradigm. (The submitted write-up is here).

Comments about the Evaluation Comments

Since serious work was demonstrated I put a lot of effort into comments below, also because a lot of surface for commenting was left exposed. Soft spots have been identified but given the breadth of models used I suspect I skimmed too quickly on simple comments that could have been added at the specific model level — I took the approach of highlighting as many key conceptual aspects at the expense of diving into the feature list of the particular MEMM/CRF (they can always specialize more in a particular model family but there were many important fundamental comments to make).

Comments While Reviewing

- Very orderly project code with pointers to the obtained result files
- There could have been an example or two in the paragraph discussing the Hebrew phonology-morphology relationship.

Comments to Task definition

- it's nice to include a non-plagiarized diagram (unlike most other projects did). However you seem (?) to miss specifically in not clearly stating a rigorous definition of the input type to the machine learning part ….. despite the effort and idiomatic writing style, the task definition is lacking a clear definition of the input to the machine learning model.

  I found that the processing pipeline preceding the machine learning model could have been better presented. It is actually the same issue, since the result of the pre-processing is the input to the ML ….. The information is there but could have been more clearly put through better planning of how to present itself.

  It also lacks motivation and reference to other competing (or easily conceivable) ways of framing the task for machine learning.

Feedback To Project Team

Part I

- **Description of annotation process / quality**
  Did you annotate the vowels over the entire MILA corpus by hand? How many different annotators participated? Did you encounter any special dilema or encounter high inter-annotator disagreement between different annotators? Anyway it was worth explicitly mentioning that you annotated it yourselves (even if you think it's not perfect; I guess one would typically think their annotation is not perfect if they haven't stopped in the middle and redone it at least once).

  If only to take credit for that and avoid this obvious question, or, for making available your annotations as a resource for future research or as a basis for improved annotation which you hint at in your suggestions for future work.

- **Terminology comment:**
  model features are not hyper-parameters (Section 3.3.2), they are *the parameters* learned by the model themselves (the learned weights which are arrived at when the model training is over). In other words the learned weights are typically called simply *parameters*. This holds for machine learning in general (deep learning included). Etymologically I think that's why they call the other stuff *hyper-parameters* ….

hyper-parameters are values which are not being learned during the training of the model, but rather ones which affect how the model trains.

- **Completeness of linguistic foundation for the work:**

    - Although marginal, In the background buildup, it should have been mentioned for theoretical completeness, that some edge cases defy the constraints which you've described about the relationship between values and consonants:
        - פילח, שילח, משלוח, תפוח .....

    - It's good to aim at people (or linguists) who do not (yet) know about Hebrew at all, in describing how and why דגש should make the consonant letter it appears in belong in both the vowel preceding and the vowel following it. Maybe even this was a decision you took for formalizing the problem for machine learning (in which case it's best to motivate this decision briefly).

    - These just make the starting point more sound and approachable to a larger audience at the same time

- I find that the Hebrew phonology-morphology relationship could have received a slightly wider exposition, even if by reference to one of the cited papers, for the same reason. For example some letters can take both vowel and consonant role (or both at the same time), and discussing this could have brought about two potential kinds of benefit:

    - Help you formalize the nature of the natural language input, possibly opening up more possibilities in how to pre/post-process the natural language input for the Machine Learning.

    - Assist the readers in understanding the way that you define the input to the Machine Learning phase, and possibly make them able to reason about the motivation for this specific definition chosen. See next comment which ties in to this.

- The writing and reasoning style which you compiled is very idiomatic for the academic NLP genre and relatively grammar/typo error free, really well done given the timeframes. **Still, d**espite opting for a diagram (Figure 1), the machine learning input-output definition remained un-optimal in how it is presented. I would say one has to read down to the last word of the write-up to be sure they got it right.

  This facet should be reworked for greater reading clarity so that it is crystal clear once the reader finished reading the first paragraph in which you present the model definition. An example (Table 1) is not enough … it should go something like: "the input are words that have undergone the following transformations …..". ***It turns out that completing this sentence turns out to be harder than it seems even after having read the entire write-up***.

  Tying back to the previous comment, why is the last "h" part of the input in the example given in Table 1? How do you rigorously define the transformation of the original word to the form that is input for the model? Conceptually I could say that the liguistic status of that "h" is somewhere between being a quiet consonant and a vowel hint — a discussion about that, or one that makes it redundant, is missing.

- I think we should frame this work as dealing with *modern* Hebrew, otherwise you could say that "h" is a מפיק which people actually vocalize. Either way this represents one small void in the framing and model

definition, which may have “sibling” voids that you could avoid or explicitly mention as “out of the scope” of the current modelling.

- Just to put in context — I do presume and understand that it was important to get moving with having a complete ML model moving instead of delving on different ways of translating the problem into an ML problem, to avoid never starting any real work. So I’m just suggesting that for an ideal perfect write-up, in the future you would re-consolidate this facet more rigorously at the stage of planning the write-up, to the work to be accessible to a larger NLP audience …

- More technically and lightly, but still related, a further rigorous analysis and definition would also have meant that trivially Section 4’s definition of the input would not conflict with that of Table 1 (is the model input words or sentences?).

- I’m sure you took some inspiration from *Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes* when studying the topic. I think in fully-fledged research write-ups its good to refer back to it in your write-up, if only as a way to help the reader ground themselves in the topic (not to copy whole paragraph pieces like other student teams have done) like you have when learning the topic.

- If you haven’t — take a look at some additional probability distribution divergence metrics. Such as https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence. Or wait till they hit you in another course.

- I think this is a false affirmation, or, you should have easily quantified it at least over the (annotated) data: “If we assume the vowels i.i.d. with equal probability”

- Related to the previous point, think that a good baseline would have involved the ways that each word appears in the training data. Intuition: if you always guess that a word stripped from its vowels is the most common form that this word shows up as in the original data, then you might get a very good score on words that show up in the training data (even if not for unseen words). Zipf’s law might help even though it is Hebrew. Well, if there is no earlier baseline predating your work, for your chosen project frame.

- You have not *motivated* why edit distance is a good metric in your discussion where you use it. Maybe it was obvious to you or inspired by one of the papers, but I found it lacking (did I miss anything?).

Part II

- As to the idea of character embeddings which you diligently pursued in parallel to considering simple one-hot encoding. (BTW why MDS for dimensionality reduction? Was it used in papers preceding the age of word/character embeddings? Just curious). Obviously it’s not mentioned whether one-hot or MDS dimensionality worked better because the RNN couldn’t beat the classical models in the project experiments. Anyway this is indeed an interesting idea in this context (well, as in most deep learning contexts), and even more interesting if you can say what kind of character embedding might work beyond just reusing some existing embedding just because it already exists. This aspect is promising since embeddings may curate information from huge un-annotated data to aid a model learn over small

annotated data like you have (and like we oftentimes have for Hebrew and other “low resource” languages).

Notwithstanding, if your annotated data is huge then this is superfluous but even then it usually speeds up the network’s training substantially (the network will require far less epochs).

- There is much room to juxtapose your results against earlier ones, or if they have framed differently to mention in passing the performance that they have attained and *briefly yet concretely* how their task is different. If the task of reconstructing/predicting vowels was never exposed in a published paper before, I think the bold statement “to the best of our knowledge we are the first to attempt/explore/show ….” should be added to avoid the doubt about why no performance comparison to prior art is included.

- (Unless I really really didn’t get your input and output definition) you have framed the machine learning model as a sequence tagging problem. It could have been a major follow-up direction to also alternatively frame the task as a sequence to sequence problem! What may be the expected pros and cons should be a good thought exercise.

- I think the relationship of your work to the task of Romanization by *Ornan and Leket-Mor (2016)* is something to clear up. More generally, the relationship between Syllabification and Romanization should have been more explicitly discussed (or untangled) in the overall, or avoided in the way that you frame the task. Currently you appear to pin down the transliteration of ambiguous cases like the letters ב, פ, כ ש as a pre-processing stage before arriving at the machine learning model, and I wonder whether that’s not a way for introducing noise just because previous works suggested Romanizing the words. So I think that untangling the Romanization task would serve you well in any followup, but if it must be had then the relationships should be made explicitly IMO in the exposition parts of the write-up.

Part III

- In case any of you decided to continue with this in any way in the future, I think it would be very implied to relate this to the domain of speech transcription (colloquially called voice-to-text) or generation (text-to-speech), or to diacritization (ניקוד). In diacritization I’d suggest looking at some normalized ניקוד where historical differences between diacritic symbols that sound the same in modern Hebrew are simply merged and we have only the small set of vowel sounds that we really have in modern spoken Hebrew the way it’s currently spoken. This will also make the task of ניקוד roughly a subset or specialization of the task of speech generation in one or two scenarios.

  In speech generation I’d suggest learning entirely different ways in which speech generation models are trained, for thinking how such different approaches relate or contrast in useful ways. Spoiler alert: in speech synthesis the recent approach is to train deep neural networks over text and speech pairs without implicit knowledge of vowels and consonants …. You can look up models like Deep Speech and others that come with research papers from Google, Facebook, Baidu etc … I think there is some course in Bar-Ilan by Dr. Keshet about voice applications in the context of machine learning.

- I think that the main contribution of your work is the intellectual foundations that you have built in code and in your minds for further research on this topic (not necessarily the same task). I did not see a performance comparison to other similar tasks included in the discussion sections, so as is I would say that the CRF results look promising as an initial stage for follow-up.

- I think the relationship to two separate layers and tasks of NLP should be examined theoretically: Part of Speech Tagging and … Word Sense Disambiguation: if you already know the word sense or POS tag, you can (metaphorically speaking as if the model is a person) reuse your syllabification from earlier instances. This could be an interesting topic for open discussion in a wide context, meaning, with and without external information (data) or pre-trained models. Maybe also morphological segmentation and morphological disambiguation.

- I think it's also a possible direction for research or application, to see how well does the trained-model learn to predict specifically for unseen words. How much does it learn compared to the intuition/guess of a native speaker regarding the pronunciation of a word first encountered. (In some ways though not originally intended, this is the opposite of the previous suggestion)

- I think in future research maybe we could drop the full Syllabification frame and focus only on the vowel prediction/restoration subset of it? Do we lose anything in doing so?

- I think that if taking the refined framing approach I have humbly suggested, then similar training data can be constructed abundantly in languages where vowels are almost-always included in the text, as in many European languages. Perhaps that can then be a good benchmark, despite some orthographic-phonemic differences (Hebrew has almost no cases where consonants change when appearing in succession, and has different cases where consonants are skipped in pronunciation, than say, English does; like the 'S' in the word *shine* and the 'P' in the word *psychology,* respectively).

- I also recommend going through this book. Maybe I should have given my spare copy to you in advance to more easily navigate the related terms and gather more intuitions about phonemes.

- See also this paper that came out in September (I haven't looked deeply). And the thesis of Eran Tomer which it cites.

Grade Justification

The work shows a lot of adherence to method, and I would almost call it an over-execution in not skipping any generation of popular model. I think it's an innovative frame, I am not aware of any earlier work taking this important frame which has been made available in the public domain. I also think it's a worthy frame (notwithstanding my comments to the de-facto frame below) in being orthogonal to end-to-end speech synthesis models, and in not falling into predicting the traditional-canonical diacritics of Hebrew which have

really lost their meaning in modern spoken Hebrew. Vowel prediction has potential applications in teaching people to learn a language and is an under-explored area, certainly in Hebrew, so I find many reasons why this topic is *of importance*.

I also find this was a tough topic. The papers about it are more scarce, oftentimes written in ways confusingly conflating different aspects of morphology and phonology, and assume acquaintance with a specialized jargon of a particular ring of researchers that have already retired. It requires a lot of patience with manual annotation and examination of outputs. I have factored this personal conviction in my grading suggestion, just a bit.