From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)?

Reut Tsarfaty Dan Bareket Stav Klein Amit Seker
Bar Ilan University, Ramat-Gan, Israel
reut.tsarfaty@biu.ac.il
{dbareket,klein.stav,aseker00}@gmail.com

Abstract

It has been exactly a decade since the first establishment of SPMRL, a research initiative unifying multiple research efforts to address the peculiar challenges of Statistical Parsing for Morphologically-Rich Languages (MRLs). Here we reflect on parsing MRLs in that decade, highlight the solutions and lessons learned for the *architectural, modeling* and *lexical* challenges in the pre-neural era, and argue that similar challenges re-emerge in neural architectures for MRLs. We then aim to offer a climax, suggesting that incorporating *symbolic* ideas proposed in SPMRL terms into nowadays *neural* architectures has the potential to push NLP for MRLs to a new level. We sketch a strategies for designing Neural Models for MRLs (NMRL), and showcase preliminary support for these strategies via investigating the task of *multi-tagging* in Hebrew, a morphologically-rich, high-fusion, language.

1 Introduction

The ability to process natural language data and to automatically extract structured meanings out of them has always been the hallmark of Artificial Intelligence (AI), and today it is also of immense practical value in downstream technological applications for Information Extraction, Text Analytics, and diverse Data Science applications. The introduction of deep learning models (Goodfellow et al., 2016) into Natural Language Processing (NLP) has led to an explosion in the Neural models and pre-training techniques applied to NLP tasks — from classical tasks as tagging and parsing to end-to-end tasks as machine translation and question answering — raising the performance bar on these tasks to an all-times peak. So far though, these advances have been reported mostly for English. Can these advances carry over to languages that are typologically vastly different from English, such as *Morphologically-Rich Languages*?

The term *Morphologically-Rich Languages* (MRLs) refers to languages such as Arabic, Hebrew, Turkish or Maltese, in which significant information is expressed morphologically, e.g., via word-level variation, rather than syntactically, e.g., via fixed word-order and periphrastic constructions, as in English. These properties lead to diverse and ambiguous structures, accompanied with huge lexica, which in turn make MRLs notoriously hard to parse (Nivre et al., 2007; Tsarfaty, 2013). A decade ago, Tsarfaty et al. (2010) put forth three overarching challenges for the MRLs research community:

> *(i) The Architectural Challenge:* What *input units* are adequate for processing MRLs?
> *(ii) The Modeling Challenge:* What *modeling assumptions* are adequate for MRLs?
> *(iii) The Lexical Challenge:* How can we cope with extreme *data sparseness* in MRLs lexica?

For NLP in the pre-neural era, effective solutions have been proposed and successfully applied to address each of these challenges for MRLs, using data from MRLs treebanks and designated shared tasks (Nivre et al., 2007; Seddah et al., 2013a, 2014a; Nivre et al., 2016). The solutions proposed to the above challenges included: *(i)* parsing *morphemes* rather than words, *(ii) joint modeling* of local morphology and global structures, and *(iii)* exploiting *external knowledge* to analyze the long tail of unattested word-forms.

Upon the introduction of Neural Network models into NLP (Goldberg, 2016), it was hoped that we could dispense with the need to model different languages differently. Curiously though, this has not been the case. Languages with rich morphology typically require careful treatment, and often the design of additional resources (cf. Czarnowska et al. (2019)). Moreover, current modeling strategies for neural NLP appear to stand *in contrast* with the pre-neural proposals for processing MRLs.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7396–7408
July 5 - 10, 2020.

First, unsupervised pre-training techniques employing language modeling objectives (LM, MLM) are applied nowadays to *raw words* rather than morphemes, and deliver *word-embeddings* agnostic to internal structure. While some morphological structure may be implicitly encoded in these vectors, the *morphemes* themselves remain un-accessible (Vania et al., 2018; Cotterell and Schütze, 2015).

Second, pre-neural models for parsing MRLs call for *joint* inference over local and global structures, tasking multiple, ambiguous, morphological analyses (a.k.a. *lattices*) as input, and disambiguating these morphological structure jointly with the parsing task (Goldberg and Tsarfaty, 2008; Green and Manning, 2010; Bohnet et al., 2013a; Seeker and Centinoglu, 2015; More et al., 2019). In contrast, pre-trained embeddings select a single vector for each input token — prior to any further analysis.

Finally, pre-trained embeddings trained on *words* cannot assign vectors to unseen words. The use of unsupervised *char-based* or *sub-word* units (Bojanowski et al., 2017) to remedy this situation shows mixed results; while these models learn *orthographic* similarities between seen and unseen words, they fail to learn the *functions* of sub-word units (Avraham and Goldberg (2017); Vania and Lopez (2017) and references therein).

This paper aims to underscore the challenges of processing MRLs, reiterate the lessons learned in the pre-neural era, and establish their relevance to MRL processing in neural terms. On the one hand, technical proposals as pre-trained embeddings, fine-tuning, and end-to-end modeling, have advanced NLP greatly. On the other hand, neural advances often overlook MRL complexities, and disregard strategies that were proven useful for MRLs in the past. We argue that breakthroughs in *Neural Models for MRLs* (NMRL) can be obtained by incorporating *symbolic* knowledge and *pre-neural* strategies into the end-to-end neural architectures.

The remainder of this paper is organized as follows. In Section 2 we survey the methodological changes that neural modeling brought into NLP. In Section 3 we characterize MRLs and qualify the challenges that they pose to neural NLP. In Section 4 we assess the compatibility of pre-neural modeling and current neural modeling practices for MRLs, and in Section 5 we suggest to re-frame pre-neural solution strategies in neural terms. In Section 6 we present preliminary empirical support for these strategies, and in Section 7 we conclude.

2 The Backdrop: From Classical Natural Language Processing to End-to-End Deep Learning

Classical NLP research has been traditionally devoted to the development of computer programs called *parsers*, that accept an utterance in a human language as input and deliver its underlying linguistic structure as output. The output may be of various sorts: *Morphological* parsing analyzes the internal structure of words. *Syntactic* parsing analyses the structure of sentences. *Semantic* parsing assigns a formal representation to the utterance, one that reflects its meaning. *Discourse* parsing identifies the discourse units, discourse relations, as well as rhetoric and pragmatic structure associated with complete narratives. Since natural language exhibits ambiguity at *all* levels of analysis, *statistical* parsers aim to learn how to pick the best analysis from multiple suitable candidates (Smith, 2011).

The introduction of Deep Learning has revolutionized all areas of Artificial Intelligence, and NLP research is no exception (Goldberg, 2016). Neural-network models now demonstrate an all-times peak in the performance of various NLP tasks, from conventional tasks in the NLP pipeline like tagging and parsing (Alberti et al., 2015; Nguyen et al., 2017; Zhou et al., 2019) to diverse downstream applications, such as machine translation (Bahdanau et al., 2014; Luong et al., 2015), question answering (Andreas et al., 2016), text-to-code generation (Hayati et al., 2018) and natural language navigation (Mei et al., 2016). In addition to revolutionizing *empirical* NLP, neural models have also altered the *methodology* of conducting NLP research, in various ways, which we review here in turn.

First, while state-of-the-art models for structure prediction in NLP used to rely heavily on intricate formal structures and carefully designed features (or feature-templates) (Zhang and Nivre, 2011; Zhang and Clark, 2011a), current neural models provide a form of representation learning and may be viewed as *automatic feature-extractors* (Kiperwasser and Goldberg, 2016; Dozat and Manning, 2018). That is, as long as the input object can be represented as a vector, the neural model will learn how to map it to the appropriate set of structural decisions, without having to write features or feature-templates by hand.

Second, most neural models for NLP rely on *pre-training*, the process of acquiring word-level vector representations termed *word-embeddings*.

These vectors are used as input, instead of actual words. Initially, word embeddings were *non-contextualized* (Mikolov et al., 2013; Pennington et al., 2014), i.e., they assigned the same vector to the occurrences of a word in different contexts. Later models present *contextualized* embeddings (Devlin et al., 2018; Peters et al., 2018; Yang et al., 2019; Liu et al., 2019b), they assign different vectors to the occurrences of the same word in different contexts. Embeddings in general, and contextualized ones in particular, dramatically increased the performance of any NLP task they were applied to.

Third, working with contextualized embeddings has been so successful, that it shifted the focus of NLP practitioners from training models from scratch to *fine-tuning* (Liu et al., 2019a) pre-trained embeddings. That is, instead of tailoring hugely complex models for specific tasks and training them from scratch, a huge effort is invested in learning a general language model (LM) that can assign contextualized embeddings to words. These vectors are often argued to capture, or *encode*, various aspects of structure and meaning (Hewitt and Manning, 2019), and then, a relatively small amount of task-specific data may be used to fine-tune the pre-trained embeddings, so that the model can solve a particular task at hand.

Finally, traditional NLP tasks, such as the parsing layers mentioned earlier, were typically organized into a pipeline turning unstructured texts gradually into more complex structures by gradually increasing the complexity of analysis. Eventually, complex semantic structures formed the basis for the design of dialogue systems, question answering systems, etc. Nowadays, NN models for complex semantic tasks are often designed and trained *end-to-end (E2E)* on examples of input-output pairs. There is an implicit assumption that all relavnt linguistic features are already *encoded* in the pre-trained representations, and that they will be *automatically extracted* in the learning process.

This methodology of *pre-training*, *automatic feature extraction* and *fine-tuning* has been applied to a wide variety of tasks and saw immense success for English — and also for similar languages. Notwithstanding, the majority of achievements and results for complex *natural language understanding* (NLU) does not yet carry over to all languages, and in particular, for languages known as *Morphologically-Rich Languages*.

3 The Challenge: NLP for Morphologically-Rich Languages

The term *Morphologically-Rich Languages* entered the NLP research community about a decade ago (Tsarfaty et al., 2010) bringing to the forefront of the research a set of languages which are typologically different from English and share a host of similar processing challenges. Subsequent SPMRL events and shared tasks (Seddah et al., 2013b; Tsarfaty, 2013; Seddah et al., 2014b) illustrated how methodologies and modeling assumptions for English NLP often break down in the face of such typologically diversity. That is, while most NLP models can *in principle* be trained on data in any given language,[1] such models are often developed with English in mind, and the *bias* injected into such models is not optimal for languages that exhibit flexible word order, and rich word-internal structure, as is the case in MRLs.

Let us briefly survey the properties of MRLs and the challenges associated with them, and observe how pre-neural studies proposed to address them.

The Essence of MRLs. The term *morphologically rich languages* (MRLs) refers to languages in which significant information regarding the units in the sentence and the relations between them is expressed morphologically, i.e., via word structure, rather than syntactically, e.g., using word order and rigid structures. Morphologically-marked information may be of various sorts. For example, consider the following Hebrew sentence:[2]

(1) *hild hpil at hspr fl hildh.*
literally: the-kid.MASC.SING cause-to-fall.MASC.PAST ACC the-book of the-kid.FEM.SING
trans: "the boy made the book of the girl fall."

There are several lessons to be learned from (1). First note that the 6 tokens in Hebrew correspond to 9 tokens in the English translation — we can observe three types of morphological phenomena that has led to this. First, elements such as prepositions, relativizers and the definite markers *h* (the) in Hebrew always attach as CLITICS to lexical hosts, and do not stand on their own. Second, features as gender, number, person, tense etc. are marked by INFLECTIONAL morphemes. In particular, the final *h*

[1] E.g., via applying them to the universal dependencies (UD) treebanks (Nivre et al., 2016).

[2] In the transliteration of Simaan et al. (Simaan et al., 2001).

distinguishes *ildh* kid.FEM from its *ild* kid.MASC counterpart. Interestingly, an initial *h* marks definiteness in *hild, hspr* and *hildh*, so there is no 1:1 relation between surface elements (chars) and what they can mark. Finally, the Hebrew verb, *hpil*, which also begins with an *h*, corresponds to the construction ("*binyan*", pattern) 'cause-to-fall' via a DERIVATIONAL morphological process that combines the pattern *h__i_* (causative) and the lexical root *n.p.l* (to fall). Note that the *h__i_* causative morpheme is non-concatenative. Moreover, when combining *h__i_* + *n.p.l* into *hpil* the *n* drops, leaving only a part of the root explicit.

This word-level complexity then requires decomposition of raw surface tokens into constituent morphemes in order to transfer them to the syntactic, semantic, or downstream tasks that require this information. However, rich morphology may lead to extreme ambiguity in the decomposition of tokens into morphemes. Take for example the two occurrences of the word form *hpil* in (2):

(2) *hild hpil at hpil.*
literally: the-kid.MASC.SING cause-to-fall.MASC.PAST ACC the-elephant
translated: "the boy made the elephant fall."

Two different morphological processes lead to two different decompositions of *hpil*, one is concatenative: "the" + "elephant" (*h+pil*) and one is not: "cause-to" + "fall" (*h__i_* + n.p.l). Moreover, neither interpretation is a-priory more likely than the other. We need the global context in order to select the single human-perceived analysis for each form.

The Typology of MRLs. The extent to which morphological phenomena is reflected in different languages varies, and linguistic typology describes morphological diversity along two dimensions. One is the *synthesis* dimension, which captures the ratio of *morphemes per word*. *Isolating* languages on one end present one-morpheme-per-word, like most words in English. At the other end we have *polysynthetic languages*, where multiple morphemes can form a single word, as it is in Turkish. The other dimension is *fusion*, and it refers to *how easy* it is to decompose the word into morphemes. In Turkish, which is *agglutinative*, the segmentation into morphemes is rather straightforward. This stands in contrast with *fusional* languages, such as Hebrew, where the decomposition of a word like *hpil* is less trivial due to the intricate 'fusion' processes that went into creation.

Key Challenges in NLP for MRLs The linguistic characteristics of MRLs are known to pose challenges to the development of NLP models, shared across languages and tasks. The overarching challenges are summerized in Tsarfaty et al. (2010):

> (i) THE ARCHITECTURAL CHALLENGE: What are the units that should enter as input into the NLP pipeline for MRLs? Are they words? Morphemes? How are these units identified and propagated down the pipeline?
> (ii) THE MODELING CHALLENGE: What are the modeling assumptions that are appropriate for models for MRLs? What kind of structure representations and features (or feature-templates) are appropriate?
> (iii) THE LEXICAL CHALLENGE: How can we cope with the extreme data sparseness that follows from the complex structure of words and the productivity of morphology?

Pre-Neural Solutions in NLP for MRLs. Let us now survey the solutions proposed for these three overarching challenges in the *pre-neural* era.

In response to the ARCHITECTURAL challenge, several input alternatives have been proposed. The input to processing an MRL can be composed of raw tokens, segmented morphemes, or complete morphological lattices that capture the multiple possible analyses for each input tokens (More et al., 2018). Morphological lattices seem particularly advantageous, since on the one hand they represent the explicit decomposition of words into morphemes, and on the other hand retain the morphological ambiguity of the input stream, to be disambiguated downstream, when information from later phases, syntactic or semantic, becomes available.

Lattice-based processing has led to re-thinking the MODELING architectures for MRLs, and to propose JOINT models, where multiple levels of information are represented during training, and are jointly predicted at inference time. Such joint models have been developed for MRLs in the context of phrase-structure parsing (Tsarfaty, 2006; Goldberg and Tsarfaty, 2008; Green and Manning, 2010) and dependency parsing (Bohnet et al., 2013b; Seeker and Çetinoğlu, 2015; More et al., 2019). In all cases, it has been shown that *joint* models obtain better results than their morphological or syntactic standalone counterparts.[3]

[3] Joint models are shown to be effective for other tasks and languages, such as parsing and NER (Finkel and Manning, 2009) or parsing and SRL (Johansson and Nugues, 2008).

Finally, the LEXICAL challenge refers to the problem of *out-of-vocabulary* items. Supervised training successfully analyzes attested forms, but fails to analyze the long tail of morphological forms in the language, not yet attested during training. Pre-neural models for MRLs thus benefit from additional *symbolic* information beyond the supervised data. It can be in the form of online dictionaries, wide-coverage lexica, or a-priori knowledge of the structure of morphological paradigms in the language (Sagot et al., 2006; Goldberg et al., 2009).

Where We're At Upon the introduction of neural models into NLP the hope was that we could dispense with the need to develop language-specific modeling strategies, and that models will seamlessly carry over from any one language (type) to another. Curiously, this was not yet shown to be the case. NLP advances in MRLs still lag behind those for English, with lower empirical results on classical tasks (Straka et al., 2016), and very scarce results for applications as question answering and natural language inference (Hu et al., 2020).

More fundamentally, NLP researchers nowadays successfully predict linguistic properties of English via neural models as in Linzen et al. (2016); Gulordava et al. (2018), but they are less successful in doing so for languages that differ from English, as in Ravfogel et al. (2018). It is high time for the MRL community to shed light on the methodological and empirical gaps between neural models for English and for MRLs, and to bridge this gap.

4 The Research Objective: NLP for MRLs in the Deep Learning Era

The point of departure of this paper is the claim that neural modeling practices employed in NLP nowadays are *suboptimal* in the face of properties of MRLs. In what follows we illuminate this claim for the four neural methodological constructs that we termed *pre-training*, *fine-tuning*, *feature-extraction* and *end-to-end modeling*.

Pre-training of word embeddings presupposes that the input to an NLP architecture consists of raw words. However, word-level embeddings may not be useful for tasks that require access to the actual morphemes. For example, for semantic tasks in MRLs, it is often better to use *morphological embeddings* of lemmas rather than words (Avraham and Goldberg, 2017). Also, dependency parsing for MRLs requires access to morphological segments, according to the UD scheme (Straka et al., 2016).

A reasonable solution might be to morphologically analyze and segment all input words *prior to* pre-training. Unfortunately, this solution does not fit the bill for MRLs either. First, current neural segmentors and taggers for MRLs are not accurate enough, and errors in the analyses propagate through the pre-training to contaminate the trained embeddings and later tasks. In the universal segmentation work of (Shao et al., 2018), for instance, neural segmentation for languages which are high on both the *synthesis* and the *fusion* index, such as Arabic and Hebrew, lags far behind. Beyond that, there is the technical matter of resources. Pre-training models as Devlin et al. (2018); Liu et al. (2019b); Yang et al. (2019) requires massive amounts of data and computing resources, and such training often takes place outside of academia. Training *morphological* embeddings rather than *word* embeddings was not taken up by any commercial partner.[4]

Next, let us turn to the notion of *fine-tuning*, widely used today in all sorts of NLP tasks, typically in conjunction with contextualized embeddings as (Devlin et al., 2018; Peters et al., 2018; Liu et al., 2019b). An argument may be advanced that *contextualized embeddings* actually encode accurate disambiguated morphological analyses in their context-based representations, and all we have to do is to *probe* these vectors and make these morphological analyses explicit. This argument is appealing, but it was never seriously tested empirically, and it is an open question whether we can successfully probe the fine-grained morphological functions from these vectors.

A possible caveat for this line of research has to do with the inner-working of contextualized representations. Most contextualized embeddings operate not on words but on *word-pieces*. A word-pieces algorithm breaks down words into subwords, and the model assigns vectors to them. The word-pieces representations are later *concatenated* or pooled together to represent complete words. It is an open question whether these word-pieces capture relevant aspects of morphology. In particular, it is unclear that the strategy of relying on chars or char-strings is adequate for encoding *non-concatenative* phenomena that go beyond simple character sequencing, such as templatic morphology, substraction, reduplication, and more (Ackerman and Malouf, 2006; Blevins, 2016).

[4] Possibly since this does not align with the business goals.

The notion of word-pieces leads us to consider the LEXICAL challenge. The suggestion to use sub-word units (chars or char n-grams) rather than words could naturally help in generalizing from seen to unseen word tokens. There is a range of sub-word units that are currently employed (chars, char-grams, BPEs (Sennrich et al., 2015)), nicely compared and contrasted by Vania and Lopez (2017). Vania and Lopez (2017); Vania et al. (2018) show that for the type of sub-word units that are currently used, standard pre-training leads to clustering words that are similar *orthographically*, and do not necessarily share their linguistic *functions*. When a downstream task requires the morphological signature (e.g., dependency parsing in (Vania et al., 2018)) this information is not recoverable from models based on sub-word units alone.

On the whole, it seems that *end-to-end modeling* for MRLs cannot completely rely on *automatic feature extraction* and dispense with the need to explicitly model morphology. It is rather the contrary. Explicit morphological analyses provide an excellent basis for successful feature extraction and accurate downstream tasks. When such analysis is missing, results for MRLs deteriorate. So, we should aim to recover morphological structures rather than ignore them, or jointly infer such information together with the downstream tasks.[5]

A different, however related, note concerning *automatic feature extraction* in MRLs has to do with the flexible or free word-order patterns that are exhibited by many MRLs. Many neural models rely on RNNs (Hochreiter and Schmidhuber, 1997) for feature extraction. These models assume complete linear ordering of the words and heavily rely on positions in the process of representation learning. Even pre-training based on *attention* and *self-attention* (Vaswani et al., 2017) assign weights to *positional* embeddings. In this sense, the bias of current neural models to encode *positions* stands in contrast with the properties of MRLs, that often show discrepancies between the linear position of words and their linguistic functions. It is an open question whether there are more adequate architectures for training (or pre-training) for more *flexible* or *free word-order* languages.

[5]Furthermore, Gonen et al. (2019) have recently shown that one needs to know the *explicit* morphological analyses in order to effectively *ignore* or neutralize certain morphemes, for instance discarding gender for reducing bias in the data.

5 Research Questions and Strategies

The Overarching Goal The purpose of the proposed research theme, which we henceforth refer to as *Neural Models for MRLs* (NMRL), is to devise modeling strategies for MRLs, for classical NLP tasks (tagging, parsing) and for downstream language understanding tasks (question answering, information extraction, NL inference, and more). This research diverges from the standard methodology of applying DL for NLP in three ways.

First, current *end-to-end* neural models for complex language understanding are developed mostly for English (Wang et al., 2018, 2019). Here we aim to situate neural modeling of natural language understanding in cross-linguistic settings (e.g., (Hu et al., 2020)). Second, while current neural models for NLP assume pre-training with massive amounts of unsupervised data (Ruder et al., 2019; Yang et al., 2019; Liu et al., 2019b), research on MRLs might be realistically faced with resource-scarce settings, and will require models that are more "green" (Schwartz et al., 2019). Finally, while many neural-based models developed for English presuppose that linguistic information relevant for the downstream task is implicitly encoded in word vectors, and may be successfully predicted by neural models (Linzen et al., 2016), we *question* the assumption that ready-made pre-trained embeddings, will indeed encode all relevant information required for end-to-end models in MRLs.

The key strategies we propose in order to address NMRL include transitionining to (i) *morphological-embeddings*, (ii) *joint lattice-based modeling*, and (iii) *paradigm cell-filling* (Blevins, 2016; Ackerman et al., 2009), as we detail shortly.

Research Questions. To instigate research on NMRL, let us define the three overarching DEEP challenges of MRLs in the spirit of (Tsarfaty et al., 2010). For these challenges, the aim is to devise solutions that respect the linguistic complexities while employing the most recent deep learning advances.

- THE DEEP ARCHITECTURAL CHALLENGE: The 'classical' architectural challenge aimed to define optimal input and output units adequate for processing MRLs. In neural terms, this challenge boils down to a question concerning the *units* that should enter pre-training. Are they words? Word-pieces? Segmented morphemes? Lemmas? Lattices? Further-

more, should these units be predicted from existing pre-trained embeddings (e.g., multilingual BERT (Ruder et al., 2019) or XLNet (Yang et al., 2019)), or should we develop new pre-training paradigms that will make the relevant morphological units more explicit?

- THE DEEP MODELING CHALLENGE: The use of neural models for NLP tasks re-opens an old debate concerning *joint* vs *pipeline* architectures for parsing MRLs. The strategy of *pre-training* word vectors and then employing *feature extraction* or *fine-tuning* pre-supposes a pipeline architecture, where a model sets all morphological decisions during *pre-training*. Joint models assume lattices that encode ambiguity and partial order, and morphological disambiguation happens only later, in the global context of the task. Is it possible to devise neural *joint models* parsing for MRLs? And if so, would they still outperform a pipeline?

- THE DEEP LEXICAL CHALLENGE: Despite the reliance on pre-trained embeddings and unsupervised data, there is still an extreme amount of unseen lexical items in the long tail of inflected forms in the language, due to the productive nature of morphology. Therefore, we need to effectively handle words outside of the *pre-trained* vocabulary. How can we cope with the extreme data sparseness in highly synthetic MRLs? Should we incorporate external resources — such as dictionaries, lexica, or knowledge of paradigm structure — and if so, how should such symbolic information be incorporated into the end-to-end neural model?

Solution Strategies. The work on NMRL may proceed along either of these four reserch avenues, each of which groups together research efforts to address a different challenge of NMRL.

- **Neural Language Modeling for MRLs.** The strategy here is to empirically examine the ability of existing pre-trained language models to encode rich word-internal structures, and to devise new alternatives for pre-training that would inject relevant biases into the language models, and make morphological information effectively learnable. This may be done by proposing better *word-pieces* algorithms, and/or devising new pre-training objectives (e.g., lattice-based) that are more appropriate for MRLs.

- **Joint Neural Models for MRLs.** The aim here is to devise neural models that parse morphologically ambiguous input words in conjunction to analyzing deeper linguistic layers, and to investigate whether these joint models work better than a pipeline – as has been the case in pre-neural models. Neural modeling of morphology may be donw jointly with, named-entity recognition, syntactic or semantic parsing, and downstream tasks as information extraction and question answering. Interleving information from all layers may be done by all at once (e.g., via Multi-Task Learning (Caruana, 1997)) or by gradually adding complexity (e.g., via Curriculum Learning (Bengio et al., 2009)).

- **Neural Applications for MRLs.** We aim to develop effective strategies for devising *end-to-end* models for complex language understanding in MRLs. To do so, the community needs high-quality benchmarks for question answering, machine reading and machine reasoning for MRLs. Initially, we need to rely on lessons learned concerning pre-training and joint modeling in the previous items, in order to devise successful architectures for solving these tasks. Moreover, developing benchmarks and annotating them both at the morphological level and for the downstream task will help to evaluate the benefits of explicit morphological modeling versus representation learning, for acquiring word-internal information needed for the downstream task.

- **Closing the Lexical Gap for MRLs.** Finally, we need to develop effective strategies for handling out-of-vocabulary (OOV) items in neural models for MRLs. Currently, the main focus of investigation lies in breaking words into pieces, to help generalize from seen to unseen word tokens. As a complementary area of investigation, a plausible direction would be to shift the focus from the *decomposition* of words into morphemes, to the *organization* of words as complete paradigms. That is, instead of relying on sub-word units, identify sets of words organized into morphological paradigms (Blevins, 2016). Rather than construct new words from observed pieces, complete unseen *paradigms* by analogy based on observed complete paradigms.

Model → / Segmentation ↓	Pre-Neural SOTA	LSTM-CRF	LSTM-CRF +Char	LSTM-CRF +FT	LSTM-CRF +Char+FT	Seq2Seq COPYNET	BERT
Oracle	-	91.46	93.2	94.6	96.03	-	95.56
Predicted	-	86.16	86.57	90.76	92.57	-	92.27
Raw Tokens	-	73.38	79.26	88.63	91.81	-	92.57
Raw Lattices	**95.5**	NA	NA	NA	NA	**95.1**	NA

Table 1: F-Scores for Hebrew Multi-tagging on the standard dev-set of the Modern Hebrew treebank. +Char means a character-based LSTM encoding, +FT means morphologically-trained Fast-text embeddings. BERT means fine-tuning the contextualized embeddings of Multilingual BERT (Ruder et al., 2019). COPYNET is the model we propose in Section 6. *Oracle* Segmentation means that the segmentation into morphemes (expert annotation) is known in advance. *Predicted* Segmentation means the decomposition into morpheme automatically predicted via More et al. (2019). *Raw Tokens* means that raw input tokens are provided as is, *Raw Lattices* means that the tokens are automatically transformed into complete morphological lattices based on a wide-coverage symbolic lexicon.

Expected Significance. As has been the case with SPMRL, work on NMRL is expected to deliver architectures and modeling strategies that can carry across MRLs, along with a family of algorithms for predicting, and benchmarks for evaluating, a range of linguistic phenomena in MRLs. From a *scientific* standpoint, this investigation will advance our understanding of what types of linguistic phenomena neural models can encode, and in what ways properties of the language should guide the choice of our neural architectural decisions. From a *technological* point of view, such modeling strategies will have vast applications in serving language technology and artificial intelligence advances to a range of languages which do not currently enjoy these technological benefits.

6 Preliminary Empirical Evidence

Goal. In this section we aim to empirically assess the ability of neural models to recover the word-internal structure of morphologically complex and highly ambiguous surface tokens in Modern Hebrew. Hebrew is a Semitic language which lies high on both the *synthesis* and *fusion* typological indices, and thus provides an interesting case study.

Specifically, we devised a *multi-tagging* task where each raw input token is tagged with the sequence of Part-of-Speech tags that represent the functions of its constituent morphemes. For example, the token *hpil* in Section 3 can assume two different *multi-tag* analyses: VERB (made-fall) or DET+NOUN (the elephant). The number of distinct tags in the *multi-tagging* analyses of Hebrew tokens can be up to seven different tags, that represent distinct functions contained in the word token.

Models. We compare the results of multi-tagging obtained by a state-of-the-art, pre-neural, morpho-syntactic parser (More et al., 2019) that is based on the structured prediction framework of Zhang and Clark (2011b).

The pre-neural parser explicitly incorporates three components for addressing the challenges associated with MRLs: (i) it receives complete *morphological lattices* as input, where each input token is initially assigned the set of all possible morphological analyses for this token, according to a wide-coverage lexicon, (ii) it employs *joint training and inference* of morphological segmentation and syntactic dependencies, and (iii) it employs unknown-words heuristics based on *linguistic rules* to assert possible valid analyses of OOV tokens.

We compared the performance of this pre-neural parser to three neural architectures:

- An end-to-end language-agnostic *LSTM-CRF* architecture, trained to predict a single complex tag (multi-tag) per token, encoding words with and without *morph/char embeddings*.
- An architecture based on the Hebrew section of *multilingual BERT*, fine-tuned to predict a single complex tag (multi-tag) per token.
- As a first approximation of incorporating symbolic morphological constructs into the neural end-to-end architecture, we designed our own COPYNET, a sequence-to-sequence pointer-network where the input consists of *complete morphological lattices* for each token, and a *copy-attention* mechanism is trained to jointly select morphological segments and tag associations *from within the lattice*, to construct the complete multi-tag analyses.

Data and Metrics. We use the Hebrew section of the SPMRL shared task (Seddah et al., 2013b) using the standard split, training on 5000 sentences and evaluating on 500 sentences. For generating the lattices we rely on a rule-based algorithm we devised on top of the wide-coverage lexicon of (Adler and Elhadad, 2006), the same lexicon employed in previous work on Hebrew (More and Tsarfaty, 2016; More et al., 2019; Tsarfaty et al., 2019). We report the F-Scores on Seg/POS as defined in More and Tsarfaty (2016); More et al. (2019).

Results. Table 1 shows the multi-tagging results for the different models. The pre-neural model obtains 95.5 F1 on joint Seg+POS prediction on the standard dev set. As for the neural models, in an *oracle* segmentation scenario, where the gold morphological segmentation is known in advance, both BERT and the LSTM-CRF get close to the pre-neural model results. However, they solve an easier and *unrealistic* task, since in realistic scenarios the gold segmentation is never known in advance. In the more *realistic* scenarios, where the segmentation is automatically predicted (via More et al. (2019)), the results of the Neural models substantially drop. As expected, morph-based and char-based representations help to improve results of the LSTM-CRF model, though not yet reaching the 95 F-score of the pre-neural model. Finally, employing our COPYNET with symbolic morphological lattices, with OOV segmentation heuristics as in the pre-neural model, leads to the most significant improvement, almost closing the gap with the pre-neural state-of-the-art result. Unfortunately, lattices are *incompatible* with LSTMs and with BERT, since LSTMs and BERT models assume complete linear ordering of the tokens, while lattices impose only a *partial order* on the morphemes. The question how to incorporate contextualized embeddings into joint, lattice-based, models is fascinating, and calls for further research.

7 Discussion and Conclusion

This paper proposes NMRL, a new (or rather, redefined) research theme aiming to develop neural models, benchmarks, and modeling strategies for MRLs. We surveyed current research practices in neural NLP, characterized the particular challenges associated with MRLs, and demonstrated that some of the neural modeling practices are incompatible with the accumulated wisdom concerning MRLs in the SPMRL literature.

We proceeded to define the three DEEP counterparts to the challenges proposed in Tsarfaty et al. (2010), namely, the DEEP ARCHITECTURAL CHALLENGE, DEEP MODELING CHALLENGE and DEEP LEXICAL CHALLENGE, and sketched plausible research avenues that the NMRL community might wish to explore towards their resolution.

Our preliminary experiments on Hebrew multi-tagging confirmed that relying on lessons learned for MRLs in the pre-neural era and incorporating similar theoretical constructs into the neural architecture indeed improves the empirical results on *multi-tagging* of Hebrew, on the very basic form of analysis of Modern Hebrew — a morphologically rich and highly-fusional language.

This type of research needs to be extended to the investigation of multiple tasks, multiple languages, and multiple possible pre-training regimes (words, chars, morphemes, lattices) in order to investigate whether this trend extends to other languages and tasks. Whether adopting solution strategies for MRLs proposed herein or devising new ones, it is high time to bring the linguistic and moprhological complexity of MRLs back to the forefront of NLP research, both for the purpose of getting a better grasp of the abilities, as well as limitations, of neural models for NLP, and towards serving the exciting NLP/AI advances to the understudied, less-privileged, languages.

Acknowledgments

We thank Clara Vania, Adam Lopez, and members of the Edinburgh-NLP seminar, Yoav Goldberg, Ido Dagan, and members of the BIU-NLP seminar, for intriguing discussions on earlier presentations of this work. This research is kindly supported by the Israel Science Foundation (ISF), grant No. 1739/16, and by the European Research Council (ERC), under the Europoean Union Horizon 2020 research and innovation programme, grant No. 677352.

References

Farrell Ackerman, James Blevins, and Robert Malouf. 2009. *Parts and wholes: Implicative patterns in inflectional paradigms*, pages 54–82.

Farrell Ackerman and Robert Malouf. 2006. Patterns of relatedness in complex morphological systems and why they matter.

Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based hmm for Hebrew morpho-

logical disambiguation. In *ACL*. The Association for Computer Linguistics.

Chris Alberti, David Weiss, Greg Coppola, and Slav Petrov. 2015. Improved transition-based parsing and tagging with neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1359, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554, San Diego, California. Association for Computational Linguistics.

Oded Avraham and Yoav Goldberg. 2017. The interplay of semantics and morphology in word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 422–426, Valencia, Spain. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Y. Bengio, Jrme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. volume 60, page 6.

James P. Blevins. 2016. *Word and Paradigm Morphology*. Oxford University Press UK.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013a. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajic. 2013b. Joint morphological and syntactic analysis for richly inflected languages. *TACL*, 1:415–428.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. *ArXiv*, abs/1909.02855.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 326–334, Boulder, Colorado. Association for Computational Linguistics.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *J. Artif. Int. Res.*, 57(1):345–420.

Yoav Goldberg and Reut Tsarfaty. 2008. A single framework for joint morphological segmentation and syntactic parsing. In *Proceedings of ACL*.

Yoav Goldberg, Reut Tsarfaty, Meni Adler, and Michael Elhadad. 2009. Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and EM-HMM-based lexical probabilities. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 327–335, Athens, Greece. Association for Computational Linguistics.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 2019 Workshop on Widening NLP*, pages 64–67, Florence, Italy. Association for Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. The MIT Press.

Spence Green and Christopher D. Manning. 2010. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of COLING*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Shirley Anugrah Hayati, Raphael Olivier, Pravalika Avvaru, Pengcheng Yin, Anthony Tomasic, and Graham Neubig. 2018. Retrieval-based neural code generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080.

Richard Johansson and Pierre Nugues. 2008. Dependency-based semantic role labeling of PropBank. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 69–78, Honolulu, Hawaii. Association for Computational Linguistics.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2772–2778. AAAI Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Amir More, Özlem Çetinoğlu, Çağrı Çöltekin, Nizar Habash, Benoît Sagot, Djamé Seddah, Dima Taji, and Reut Tsarfaty. 2018. Conll-ul: Universal morphological lattices for universal dependency parsing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Amir More, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint transition-based models for morpho-syntactic parsing: Parsing strategies for mrls and a case study from modern hebrew. In *Transactions of ACL*.

Amir More and Reut Tsarfaty. 2016. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING*, pages 337–348. The COLING 2016 Organizing Committee.

Dat Quoc Nguyen, Mark Dras, and Mark Johnson. 2017. A novel neural network model for joint POS tagging and graph-based dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 134–142, Vancouver, Canada. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke

Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Shauli Ravfogel, Francis M Tyers, and Yoav Goldberg. 2018. Can lstm learn to capture agreement? the case of basque. *arXiv preprint arXiv:1809.04022*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. 2006. The lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai. *ArXiv*, abs/1907.10597.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014a. Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. pages 103–109.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014b. Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland. Dublin City University.

Djame Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, D. Jinho Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Villemonte Eric de la Clergerie. 2013a. Proceedings of the fourth workshop on statistical parsing of morphologically-rich languages. pages 146–182. Association for Computational Linguistics.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013b. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA. Association for Computational Linguistics.

Wolfgang Seeker and Ozlem Centinoglu. 2015. A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Wolfgang Seeker and Özlem Çetinoğlu. 2015. A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Yan Shao, Christian Hardmeier, and Joakim Nivre. 2018. Universal word segmentation: Implementation and interpretation. *Transactions of the Association for Computational Linguistics*, 6:421–435.

Khalil Simaan, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a tree-bank of modern hebrew text. *Traitement Automatique des Langues*, 42(2):247–380.

Noah A. Smith. 2011. *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Reut Tsarfaty. 2006. Integrated morphological and syntactic disambiguation for modern Hebrew. In *Proceedings ACL-CoLing Student Research Workshop*, pages 49–54, Stroudsburg, PA, USA. ACL.

Reut Tsarfaty. 2013. A unified morphosyntactic scheme for stanford dependencies. In *Proceedings of ACL*.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (spmrl): What, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, SPMRL '10, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reut Tsarfaty, Amit Seker, Shoval Sadde, and Stav Klein. 2019. What's wrong with hebrew nlp? and how to make it right. In *Proceedings of EMNLP*.

Clara Vania, Andreas Grivas, and Adam Lopez. 2018. What do character-level models learn about morphology? the case of dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2573–2583,

Brussels, Belgium. Association for Computational Linguistics.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Yue Zhang and Stephen Clark. 2011a. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Yue Zhang and Stephen Clark. 2011b. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA. Association for Computational Linguistics.

Junru Zhou, Zuchao Li, and Hai Zhao. 2019. Parsing all: Syntax and semantics, dependencies and spans.