MGAD: Multilingual Generation of Analogy Datasets

Mostafa Abdou[∗⋆◇†], **Artur Kulmizev**[∗⋆♣], **Vinit Ravishankar**[†⋆△]
[∗]University of Groningen, [†]Charles University in Prague
[∗]Center for Language and Cognition, [†]Institute of Formal and Applied Linguistics
[◇]m.abdou@student.rug.nl, [♣]a.kulmizev@student.rug.nl, [△]vinit.ravishankar@gmail.com

Abstract

We present a novel, minimally supervised method of generating word embedding evaluation datasets for a large number of languages. Our approach utilizes existing dependency treebanks and parsers in order to create language-specific syntactic analogy datasets that do not rely on translation or human annotation. As part of our work, we offer syntactic analogy datasets for three previously unexplored languages: Arabic, Hindi, and Russian. We further present an evaluation of three popular word embedding algorithms (Word2Vec, GloVe, LexVec) against these datasets and explore how the performance of each word embedding algorithm varies between several syntactic categories.

Keywords: word embeddings, evaluation, multilingual, natural language processing

1. Introduction

Since the emergence of dense word embeddings (Mikolov et al., 2013c), a sizable amount of work concerning the evaluation of their quality as linguistic representations has emerged. These evaluation methods can be divided into two classes: extrinsic and intrinsic. The former involves testing the performance of embeddings on various NLP tasks such as part of speech (POS) Tagging, Machine Translation, etc. (Schnabel et al., 2015). However, these methods are computationally expensive and task-specific, i.e. they solely demonstrate whether an embedding is suitable for one particular task and are therefore not suitable as a (quick) test of general quality. Intrinsic methods, on the other hand, are designed to test the degree to which a set of embeddings can model a certain linguistic property. This typically involves constructing human evaluated datasets to directly test syntactic or semantic relationships between words (Gladkova and Drozd, 2016). Embeddings are then typically evaluated by an aggregate score (e.g. a correlation coefficient) using of a set of query words and semantically related target words. This score serves as a measure of quality.

The majority of datasets created for intrinsic evaluation have focused on non-specific word relatedness (Bruni et al., 2014) or word similarity (Hill et al., 2014). More recently, the analogy-based task first proposed by Mikolov et al. (2013a) has gained popularity. This task involves retrieving a term by solving analogy questions of the form "a is to b as c is to X" using vector arithmetic (the most recognisable example of which is: *king - man + woman = queen*). However, the vast majority of analogy datasets have been constructed only for the English language, making it harder to evaluate whether the reported performance of popular embedding algorithms continue to hold when evaluated against previously untested languages. As such, this paper is concerned with generating multilingual syntactic analogy datasets for intrinsic evaluation of word embeddings. Specifically, we present a method of leveraging existing resources such as dependency parsed corpora to automatically generate the datasets. We argue that this method is advantageous in multilingual environments, where inflectional morphology can vary greatly between languages.

In Section 2., we briefly review the relevant literature in the domain. Section 3. provides more detail about what we are trying to achieve with our system, and why this is necessary. Section 4. describes our methodology in detail. Section 5. describes our evaluation parameters and how they are meaningful, whilst Section 6. describes our actual results, and demonstrates an analysis from a quantitative and qualitative perspective. Finally, we describe potential future extensions and improvements in Section 7. and conclude in Section 8.

2. Related Work

The Google analogy dataset (Mikolov et al., 2013b) is arguably the most widely adopted analogy dataset, comprising of 10,675 syntactic questions and 8,869 semantic questions (19,544 total). The former category consists of questions that aim to capture syntactic regularities that manifest in English (e.g. `good : better :: rough : ?`). Such questions were generated by POS-tagging a corpus of 267M words and extracting varying forms of adjectives, nouns, and verbs as represented by the Penn TreeBank (Marcus et al., 1993)(`JJ, JJR, JJS, NN, NNS`, etc.). The semantic questions portion of the dataset is largely comprised of `capital-country` pairs, which were manually chosen by a group of annotators. These pairs were then randomly combined with other such pairs in order to produce the entirety of the semantic portion.

Gladkova et al. (2016) expanded on the work of Mikolov et al. (2013b) in generating a balanced set of 99,200 analogy questions titled "The Bigger Analogy Test Set" (hereafter referred to as 'BATS'). These analogies were again separated into syntactic and semantic categories, each of which were further partitioned into inflectional/derivational and lexicographic/encyclopedic questions, respectively. In total, the project covered 40 different linguistic relations, each of which were represented by 2,480 questions. A distinguishing feature of BATS, outside of its breadth, was that it was designed to reduce homonymy, which occured often in

[⋆]Equal contribution. Listing order is random.

the Google analogy set. This was done by removing every word in the reference corpus that was attributed to more than one part of speech in the English WordNet. As such, pairs like *walk, walks*, both words of which could be attributed to nouns or verbs, were excluded.
The other task employed in the intrinsic evaluation of word embeddings is the word similarity/relatedness task. A commonly utilized dataset of this type is the WordSimilarity-353 Test Collection (Agirre et al., 2009). This dataset is separated into two parts which contain 153 and 200 word pairs, respectively. Each word pair is also coupled with human judgments of the two words' relatedness, represented by a 10-point Likert scale, with 0 representing completely unrelated words and 10 representing very related or identical words. Evaluating word embeddings against this dataset typically involves finding the cosine distance between both words in a word pair and generating the Spearman correlation between the human relatedness judgments and the cosine distance.

3. Motivation

Though each of the aforementioned datasets is generally suitable for the evaluation of English word embeddings, each of them fails to introduce a sustainable framework for generating similar datasets for other languages. For example, in order to recreate the WordSimilarity-353 dataset for any other language, it would be necessary to solicit the judgments of native-speakers of that language, which is a generally expensive task. The same can be said for the semantic portions of the Google and BATS datasets, the relations of which were produced by native English speakers and might not necessarily hold cross-linguistically. Furthermore, Google and BATS' reliance on the Penn TreeBank results in a failure to capture many linguistic features that do not meaningfully occur in English (case, animacy, etc.). Therefore, this makes a direct translation of either dataset not a particularly robust approach for the evaluation of non-English embeddings. In this work, then, we propose a method for the generation of analogy datasets which can be generalized for many languages. Our approach for generating syntactic analogies demonstrates that it is possible to create datasets cross-linguistically in a manner that is both low-resource and sensitive to the particularities of any given language. We test this method on three languages: Arabic, Russian, and Hindi. Our choice of languages is based on two factors. First, the dissimilarity of the languages is meant to demonstrate the cross-lingual robustness of the method. Secondly, each of this paper's authors is a native speaker of one of the languages, and so can serve as an annotator to manually check the quality of the generated datasets.

4. Methodology

4.1. Dependencies

Constituency grammars (Tesnière, 1965) have long been employed for formal representations of language grammars. However, though relatively recent, dependency grammars are an extension of this approach that offer several advantages to the former. In particular, dependency grammars' independence from a static word order allows a number of typologically varying languages to be properly represented. These advantages are reflected in the Universal Dependencies (UD) project (Nivre et al., 2016), which attempts to model dependency relations *cross-linguistically* based on a framework that holds true across all represented languages. UD treebanks typically use the 10-column CoNLL-U format for representing sentences and dependencies. Of interest to us, however, is the features column, which stores morphological information. Similar to the dependency annotations in the schema, these features are represented by cross-linguistic labels and the same format for every language, significantly simplifying querying different languages.

4.2. Templating

Though it would be ideal to have a fully automatic method of generating morphological analogies, this fails for several reasons. The most obvious of these is that morphological distinctions used across languages are far from universal, even within language families, let alone across. Another issue is the existence of morphological information that needs to be 'controlled' for; for instance, whilst BATS includes an example analogy of the English third-person singular to the infinitive, transferring this to Hindi would be non-trivial: several 'new' features that do not exist in English would need to be controlled for, such as gender or aspect; not fixing or registering these would result in multiple analogies.
To circumvent these issues, we design a simple Python script[1] that can parse 'templates' that define precisely what analogies need to be generated for a particular language. Table 1 is a (much truncated) example of one such template; the syntax for referencing the appropriate morphological features is similar to the syntax followed by UD, allowing someone with moderate familiarity with the schema to rapidly create their own templates.
The fact that UD treebanks typically store lemmas for every word results in two classes of analogies: 'core' analogies, that include the lemma, and 'composite' analogies between two non-lemmatic categories. In English, the former category would translate to something similar to `eat:eats` while the latter would be `eats:eaten`. The core analogies are trivial to generate using the lemma field; the latter, however, involve using the lemma as a link between the two word forms participating in the analogy. The analogies are further divided into `nominal` and `verbal` categories, representing noun-based and verb-based analogies, respectively. Table 2 displays the number of analogies generated, along with the distribution of their class and syntactic categories. The first row, **Type**, shows the number of 'core' and 'composite' analogies; **POS** shows the number of nominal and verbal analogies; and the last row shows the total number of analogies. Table 3 shows a selection of the generated analogy templates along with examples in each of the three languages.

4.3. Corpora

Whilst some treebanks such as Czech and Russian are large enough to provide a sufficient amount of lexical entries to

[1] All code, datasets, and templates are publicly available at: https://github.com/rutrastone/MGAD

NOUN|Number=Plur|Case=Nom	NOUN|Number=Sing|Case=Nom
NOUN|Number=Plur|Case=Acc	NOUN|Number=Sing|Case=Acc
VERB|Aspect=Perf|Gender=Masc|Number=Sing|VerbForm=Part	VERB|Case=Nom|VerbForm=Inf
VERB|VerbForm=Conv	VERB|Case=Nom|VerbForm=Inf
VERB|Case=Acc|VerbForm=Inf	VERB|Case=Nom|VerbForm=Inf

Table 1: Truncated example of a template.

	Russian	**Arabic**	**Hindi**
Type	$^{12.5}/_{35}$	$^{7.5}/_{12.5}$	$^{15}/_{12.5}$
POS	$^{25}/_{22.5}$	$^{2.5}/_{17.5}$	$^{10}/_{17.5}$
=	47.5	20	27.5

Table 2: Number of analogies (in 1,000s) generated per Type (core/compound), Part of Speech (noun/verb), and total number of analogies per language.

generate an analogy set, many are often too small and too domain-constrained to provide enough data. Therefore, we use a larger morphologically analysed corpus (Wikipedia dumps) to generate our data set. We use the MorphoDita (Straková et al., 2014) morphological analyser, trained on the manually annotated treebanks. The dependency parsing pipeline UDPipe (Straka and Straková, 2017) conveniently provides an easy-to-use wrapper for MorphoDita.

From these parsed corpora, we use our template extraction script to extract all relevant linguistic information. In generating analogies for Russian nouns, for example, we build the nominal part of our analogy set with all noun case combinations, with fixed number (eg. `+Nom+Sg:+Gen+Sg`, and with varying number and fixed case (eg. `+Nom+Pl:+Gen+Pl`. This has the effect of covering the entire breadth of cases and numbers that can possibly exist in Russian and is thus a robust evaluation of how these features are represented in any target embedding space.

An issue with this method is that languages with richer morphology would, undoubtedly, generate a much larger number of analogies than more morphologically simpler languages. We posit that this side-effect is a beneficial one, allowing the output dataset to fully assess the morphological breadth of inflectionally-rich languages on which the target embeddings are trained.

The second issue in generating our dataset is the question of word pair frequency. We follow the method used in BATS, in selecting the fifty most frequent pairs of each relation in our corpora. The frequency of a composite pair is, however, set to the minimum of the frequencies of its two constituent core pairs.

5. Evaluation

5.1. Word embeddings

In order to evaluate our datasets, we train word embeddings for each of the languages using three popular unsupervised methods: Skipgram with negative sampling (SGNS)(Mikolov et al., 2013c), GloVe (Pennington et al., 2014), and LexVec (Salle et al., 2016). These methods are a small but representative subset of the vast number of algorithms which have been proposed. However, as our focus is on evaluation datasets, we restrict our testing to these three methods and make no attempt to fine-tune the embeddings. All sets of embeddings are of dimension 300 and were trained with a window-size 5 for 5 epochs. SGNS and LexVec embeddings were trained with a learning rate of 0.05 and with 5 negative samples per training iteration, while GloVe embeddings were trained with the default learning rate of 0.025. All embeddings were trained with a minimum word count of 10 as the threshold for inclusion in the vocabulary. The Russian and Arabic embeddings were trained on their respective wikipedia dumps. However, HindMonoCorp 0.5 (Bojar et al., 2014) was used instead for Hindi, as the wikipedia corpus was significantly smaller than the others.

5.2. The Task

The task is to retrieve an answer to the question "a is to b as c is to X" as represented by hidden vector d, which is calculated as $argmax_{d \in V}(sim(d, c - a + b))$ where V is the vocabulary excluding vectors a, b, and c. We define similarity as the angular distance (cosine similarity) between vectors u and v:

$$similarity(w_1, w_2) = \frac{\vec{w_1} \cdot \vec{w_2}}{\|\vec{w_1}\| \|\vec{w_2}\|}$$

Though work by Levy and Goldberg (2014) and Linzen (2016) has shown that other functions may outperform cosine similarity on the analogy task, we nonetheless employ it in the interest of comparability with the majority of previous work.

6. Results

We report coverage[2] and accuracy over the full test datasets and separately for the nominal and verbal categories for each language per set of embeddings. A question is not covered if one or more of the words contained in it are not found in the embedding's vocabulary. Table 4 is a summary of the results.

6.1. Quantitative

Table 4 demonstrates that each embeddings model performs similarly across all three languages. This indicates the cross-linguistic stability of the datasets as a method of evaluation. Several things stand out in the results, however - particularly Hindi's excellent coverage. This is likely because the domain of our corpus is very similar to the domain of the dependency treebank, and also possibly significantly cleaner than Wikipedia as a corpus. Also, it is apparent that GloVe fares comparatively much worse for Russian

[2]Slightly different coverage over a language's dataset for different sets of embeddings is explained by the 'minimum word count' parameter being considered exclusive or inclusive (GloVe embeddings have a slightly larger vocabulary).

<table>
<tr><th>Category</th><th>Template Rule</th><th>Example</th><th>Gloss</th></tr>
<tr><td rowspan="6">Nominal</td><td><code>NOUN|Number=Plur|Case=Nom :</code></td><td>महीने महीना</td><td>month.PL month.SG</td></tr>
<tr><td><code>NOUN|Number=Sing|Case=Nom</code></td><td>मामले मामला</td><td>issue.PL issue.SG</td></tr>
<tr><td><code>NOUN|Number=Plur|Case=Nom|Definite=Ind :</code></td><td>مسؤولون مسئول</td><td>official.PL official.SG</td></tr>
<tr><td><code>NOUN|Number=Sing|Case=Nom|Definite=Ind</code></td><td>مصادر مصدر</td><td>source.PL source.SG</td></tr>
<tr><td><code>NOUN|Number=Sing|Case=Dat :</code></td><td>человеку человек</td><td>person.DAT person.NOM</td></tr>
<tr><td><code>NOUN|Number=Sing|Case=Nom</code></td><td>жизни жизнь</td><td>life.DAT life.NOM</td></tr>
<tr><td rowspan="6">Verbal</td><td><code>VERB|Aspect=Perf|Gender=Masc|Number=Sing|VerbForm=Part :</code></td><td>दिखाया दिखाना</td><td>show.PERF.3MSG show.INF</td></tr>
<tr><td><code>VERB|Case=Nom|VerbForm=Inf</code></td><td>कहा कहना</td><td>say.PERF.3MSG say.INF</td></tr>
<tr><td><code>VERB|Aspect=Imp|Gender=Masc|Person=3|Number=Sing :</code></td><td>يريد أراد</td><td>want.IMPF.3MSG want.PERF.3MSG</td></tr>
<tr><td><code>VERB|Aspect=Perf|Gender=Masc|Person=3|Number=Sing</code></td><td>يتوجه توجه</td><td>head_to.IMPF.3MSG head_to.PERF.3MSG</td></tr>
<tr><td><code>VERB|Person=3|Number=Sing|Tense=Pres :</code></td><td>живет жить</td><td>live.PRES.3SG live.INF</td></tr>
<tr><td><code>VERB|Aspect=Imp|VerbForm=Inf</code></td><td>стоит стоять</td><td>stand.PRES.3SG stand.INF</td></tr>
</table>

Table 3: Generated analogy templates and corresponding example analogies in Hindi, Arabic, and Russian

		word2vec		GloVe		LexVec	
		Coverage	Accuracy	Coverage	Accuracy	Coverage	Accuracy
Russian	**N**	79.09	25.41	81.78	17.91	79.09	22.23
	V	28.70	36.75	81.17	18.75	28.70	33.43
	=	55.23	28.20	81.50	18.30	55.23	25.09
Arabic	**N**	88.52	33.19	97.98	13.54	88.52	30.14
	V	56.30	24.69	98.67	34.41	56.30	22.93
	=	64.36	27.61	98.50	29.21	64.36	25.41
Hindi	**N**	98.02	33.14	99.98	62.85	98.02	29.45
	V	86.41	40.65	87.94	25.01	86.41	38.38
	=	90.62	37.69	92.32	39.91	90.62	34.87

Table 4: Word embedding performance on three generated sets. N, V and '=' indicate performance on nominal and verbal sections, and on a combination of the two

than it does for Arabic and Hindi. This could be related to the model's count-based implementation (as opposed to the prediction-based word2vec and LexVec), which may fail to represent low-occurring case inflection contexts in case-rich languages such as Russian.

6.2. Qualitative

Each of the three generated datasets was manually checked for correctness by one of the three authors who is a native speaker of the language. Whilst judging the correctness of our datasets is trivial, there is no simple method to judge their validity. The distinction here is that our choice of morphological forms for analogies is partly arbitrary, motivated solely by linguistic intuition of the language. There were also some areas where we were limited by the treebanks themselves: for instance, it was impossible to include inflectional adjective degrees (such as the comparative and superlative), since UD does not normalise lemmas across adjective degrees, making it impossible to "link" a positive adjective with its equivalent superlative form. For Arabic, case could not be included as a category due to both the Wikipedia corpus and the UD treebank not featuring diacritics which are used to distinguish the case of a noun.

6.3. Validity

In order to further ascertain the validity of our framework, we generate an English analogy dataset and use it to evaluate three sets of commonly utilized, publicly available pre-trained word embeddings: the Google News corpus embeddings (Mikolov et al., 2013a), Glove embeddings trained on Common Crawl (Pennington et al., 2014) (42B tokens), and LexVec embeddings trained on Common Crawl (58B tokens). We then compare the results to evaluation using the the syntactic half of the BATS dataset. As mentioned earlier, the syntactic portion of BATS includes two subcategories: inflectional and derivational analogies. The former class consists of structures such as regular plurals (`student:students`), infinitive:participle (`follow:following`) and participle:past (`following:followed`). The latter is is further divided into *stem change* and *no stem change* analogies: noun+less (`life:lifeless`) VS. verb+ation (`continue:continuation`). It is important to note that, though we attempted to replicate every relation used in the set of syntactic pairs in BATS, we were unable to generate adjective degree sets; as mentioned earlier, UD does not typically map different adjective degrees to the same lemma. We therefore generated solely the nominal and verbal sets, resulting in 17,500 pairs. The exact relations used are mentioned in the original paper (Straková et al., 2014); we do not replicate them here for brevity. The results (accuracy) are shown in Table 5 and they show a strong correlation between the results obtained by evaluating using the MGAD dataset and those obtained by using the syntactic half of BATS.

	word2vec	GloVe	LexVec
BATS (Syntactic)	68.41	63.17	68.41
MGAD	69.92	68.10	64.41

Table 5: Results (accuracy) of evaluating three sets of word embeddings using the syntactic half of BATS and the dataset generated using MGAD.

7. Future Work

7.1. Semantic relations

Our method only took syntactic relations into consideration when generating the datasets. Further work will explore resources like multilingual wordnets and Wiktionary in order to generate semantic datasets to serve as a complement to the syntactic ones we've generated here. Given the breadth of languages that are annotated in the latter source, we would like to explore the extent to which semantic relations like hypernymity and meronymity could be generated in an automatic or semi-supervised manner. Furthermore, it will be important to explore how previously-annotated English-language datasets could be translated in an efficient manner while retaining uniform syntactic relations between the terms in the analogy. Though we attempted to automatically translate the capital:country relations in the Google dataset, (e.g. case inconsistently was inevitable in every language we evaluated (`Токио(nom):Япониа(nom) :: Париже(prep):Францииа(nom); Tokyo:Japan :: (in)Paris:France`).

7.2. Data sources

Our justification for using Universal Dependencies treebanks as our source of morphological data was obvious: the treebanks were cross-linguistic, and most were well-annotated morphologically. There are, however, several directions into which we could branch for more comprehensive resources, combined with our present approach. The Apertium project (Forcada et al., 2011) includes approximately 73 morphological analysers (with varying quality). These include several analysers for several minority languages that lack dependency treebanks, such as Kyrgyz (Washington et al., 2012), Marathi (Ravishankar and Tyers, 2017) and Sardinian (Tyers et al., 2017). The relative lack of standardization across Apertium morphological analysers, however, makes it significantly harder to partially automate than Universal Dependencies, and extraction would entail attaining some familiarity with a specific analyzer's structure.

8. Conclusion

In this work we presented a method for generating analogy datasets to be used for the evaluation of word embeddings for a large number of languages. Our method utilizes readily available resources (UD Treebanks and morphological analyses) and requires minimal human supervision. A quantitative evaluation reveals that our results reflect the expected performance of several popular word embedding models across the three represented languages, emphasizing the validity of our approach. Lastly, we release three syntactic analogy datasets for Russian, Arabic, and Hindi.

9. Acknowledgements

Work on this paper has been supported by the Erasmus Mundus Language and Communication Technologies grant. We thank Rik van Noord for his helpful comments.

10. Bibliographical References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014). HindMonoCorp 0.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. 11.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Gladkova, A. and Drozd, A. (2016). Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Hill, F., Reichart, R., and Korhonen, A. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. *arXiv preprint arXiv:1606.07736*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., and others. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ravishankar, V. and Tyers, F. M. (2017). Finite-state morphological analysis for marathi. In *Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing (FSMNLP 2017)*, pages 50–55.

Salle, A., Idiart, M., and Villavicencio, A. (2016). Matrix factorization using window sampling and negative sampling for improved word representations. *arXiv preprint arXiv:1606.00819*.

Schnabel, T., Labutov, I., Mimno, D. M., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *EMNLP*, pages 298–307.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.

Tesnière, L. (1965). Eléments de syntaxe structurale.

Tyers, F. M., i Font, H. A., Fronteddu, G., and Martín-Mor, A. (2017). Rule-based machine translation for the italian–sardinian language pair. *The Prague Bulletin of Mathematical Linguistics*, 108(1):221–232.

Washington, J., Ipasov, M., and Tyers, F. (2012). A finite-state morphological transducer for kyrgyz. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).