Wiktionary Normalization of Translations and Morphological Information

Winston Wu and David Yarowsky
Department of Computer Science
Center for Language and Speech Processing
Johns Hopkins University
{wswu,yarowsky}@jhu.edu

Abstract

We extend the Yawipa Wiktionary Parser (Wu and Yarowsky, 2020) to extract and normalize translations from etymology glosses, and morphological form-of relations, resulting in 300K unique translations and over 4 million instances of 168 annotated morphological relations. We propose a method to identify typos in translation annotations. Using the extracted morphological data, we develop multilingual neural models for predicting three types of word formation—clipping, contraction, and eye dialect—and improve upon a standard attention baseline by using copy attention.

1 Introduction

Wiktionary is a large, free multilingual dictionary with a wealth of information. Yawipa (Wu and Yarowsky, 2020), henceforth W&Y, is a recent Wiktionary parser billed as “comprehensive and extensible.” It has the ability to extract numerous types information from Wiktionary, including pronunciations, part of speech, translations, etymology, and a wide range of word relations, and normalize it into an easy to process tabular format. In particular, one of Yawipa’s innovations over existing parsers was extracting translations from the definition section of a dictionary definition. Confirming its easy extensibility and improving upon its comprehensiveness, we extend Yawipa’s extraction and normalization of Wiktionary in two directions: we extract translations from an unusual source, etymology glosses, and we extract morphological relations as annotated by form-of relations. This results in an addition of 282,092 new unique translations and 4,027,201 extracted morphological relations (from the 2020-04 English Wiktionary XML dump). We present an analysis that enables us to find typos in translation annotations. Using the extracted morphological data, we experiment with several new low-resource (1.5K instances) multilingual prediction tasks on clipping, contraction, and eye dialect. Our experiments with neural sequence-to-sequence models show that using copy attention can improve performance by up to 52% over a model with a standard attention mechanism.

2 Related Work

Though Wiktionary has existed since 2002, only until very recently has there been a surge of interest in using Wiktionary. Navarro et al. (2009) was one of the first to examine Wiktionary as a resource for NLP. This paper builds upon Yawipa (Wu and Yarowsky, 2020), an open-source, extensible Wiktionary parsing framework written in Julia with support for parsing a wide variety of data from multiple language editions of Wiktionary into a structured machine-readable format. Yawipa’s goal is to be comprehensive and extensible. To that end, Yawipa goes beyond existing parsers in extracting and normalizing information, such as etymology and translations, that exist outside of structured Wiktionary markup (we further this goal in this paper), and it facilitates the creation of new parsers for other Wiktionary editions. In the literature, there are similar Wiktionary parsing efforts (e.g. knoWitiary (Nastase and Strapparava, 2015), DBnary (Sérasset, 2015), and ENGLAWI (Sajous et al., 2020)), but with different goals and coverage.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.
Most studies on translation extraction have utilized the translation section of an entry: Ács (2014) using a triangulation approach, Kirov et al. (2016) for morphophological analysis, and Wu and Yarowsky (2020) as part of a comprehensive Wiktionary parsing effort. DBnary (Sérasset, 2015) is a similar effort at parsing certain lexical data, including translations, from Wiktionary into a structured format.

Regarding extracting morphological relations between words, the foremost effort is UniMorph (Kirov et al., 2016; Kirov et al., 2018; McCarthy et al., 2020), a large broad-coverage resource comprising morphological paradigms of nouns, adjectives, and verbs in 118 languages extracted from Wiktionary. Other large-scale parsing efforts for targeted tasks include NULEX (McFate and Forbus, 2011) for parsing, IWNLP (Liebeck and Conrad, 2015) for lemmatization, and WikiPron (Lee et al., 2020) for pronunciations.

Related to the word formation mechanisms we examine, Kulkarni and Wang (2018) examine word formation in slang, specifically blends, clippings, and reduplication, and Brooke et al. (2011) predict clipping using a LSA-based approach. Constructions are not typically studied in a predictive context; Volk and Sennrich (2011) disambiguates contractions as a preprocessing step in machine translation. Researchers have recently examined eye dialect in the context of spelling correction (Eryani et al., 2020; Himoro and Pareja-Lora, 2020), but to our knowledge, this paper is the first study on eye dialect generation.

3 Extracting Translations from Etymology Glosses

Wiktionary contains translations in a specialized Translation section. W&Y extract these translations, as well as “translations” from the definition section of non-English word entries. Since non-English words have English definitions (in the English Wiktionary), short definitions can be regarded as viable translations. One unusual but particularly fruitful source of translations that has not been previously considered is glosses in the Etymology section of an entry. For example, in Wiktionary the etymology of the German word *Marienkäfer* ‘ladybug’ is:

From *Maria* (given name) + *Käfer* (“beetle”).

Glosses of each component of the compound word are given in parentheses; these are the translations that we extract. The provided glosses can help disambiguate the word in cases where a word may have multiple senses (e.g. *Käfer* can refer to a beetle, a wench, or the Volkswagen car).

The decomposition of *Marienkäfer* in the above etymology entry is encoded in MediaWiki markup as `{{compound|de|Maria|pos1=given name|Käfer|t2=beetle}}`. This is a Wiktionary template with arguments separated by pipes, indicating (1) the word is a compound, (2) it is a German word, (3) the 1st component is *Maria*, (4) the part of speech of the 1st component is “given name”, (5) the 2nd component is *Käfer*, and (6) the translation of the 2nd component is “beetle”. From this example, we would extract and normalize the second component’s translation to augment the translations already extracted by Yawipa from other sources.

### Analysis

Table 1 summarizes the number of additional translations added using these etymology glosses. In short, parsing and normalizing etymology glosses results in over 282K new unique translations (a 5.9% increase) not captured by the Translations and Definitions sections processed by W&Y.

| Source                | Extracted | Unique Translations | Unique Additions |
|-----------------------|-----------|---------------------|------------------|
| W&Y Translations      | 2,379,921 | 2,165,343           | 2,165,343        |
| W&Y Definitions       | 3,025,434 | 2,953,861           | +2,335,125       |
| Our Etymology Glosses | 464,955   | 336,696             | +282,092         |
| **Total**             | 5,894,207 | 5,455,900           | 4,782,560        |

Table 1: Counts of translations extracted from Wiktionary.

The top 5 languages we extract translations from are Latin, Greek, and Proto Indo-European (common ancestor languages) and Finnish and German (highly compositional languages). We also examine specifically where in the etymology template the gloss occurs (Table 2), whether as a named argument (e.g.
t2=beetle) or as a positional (non-named) argument (e.g. {{m|la|ab||from, away from}}),
and denoted as (none) in Table 2).

|        |                |                |                |                |            |
|--------|----------------|----------------|----------------|----------------|------------|
|        | (none)         | t3 4,450       | t7 20          | gloss6 2       |            |
| t1     | 74,792         | gloss3 738     | t8 11          | gloss11 1      |            |
| t2     | 56,452         | t4 476         | gloss5 9       | t22 1          |            |
| t      | 55,376         | t5 117         | t9 3           |                |            |
| gloss1 | 23,213         | t6 53          | t11 3          |                |            |
| gloss2 | 14,084         | gloss4 28      | t10 3          |                |            |

Table 2: Histogram of argument names of etymology translations and their counts.

We find that the large majority of etymology glosses are annotated through positional arguments, indicating that the word is not a compound word. Following this, we see a large number of t1 and t2 arguments, which occur in compositional words such as compounds and affixal words (e.g. {{compound|de|Zeit|t1=time|Geist|t2=spirit}}). Note that glosses are by no means required and are often left out for compound words (e.g. {{compound|en|light|house}}). We observe some inconsistency in whether to use t or gloss; gloss seems to be the older standard, while t is the accepted convention. The larger argument numbers in this histogram also give an indication of the number of compound words and phrases and their components contained in Wiktionary.

**Typos** This analysis also allows us to automatically identify potential annotation typos (Table 3). For example, the template argument t11 in Table 2 indicates a translation of the 11th component in a compound word or phrase. The three entries with a t11 are the Dutch stokhaver, Latin aequabilis, and Hungarian amit nyer a réven, elveszti a vámón. By examining unlikely template arguments, and then verifying the presence of previous arguments (t1 through t10) we can automatically identify typos by annotators (who probably accidentally pressed the 1 key twice, since 11-part compound words are highly unlikely). Typos are then recommended to the user, who can manually correct the upstream source.

| Lang | Word          | Etymology Template                                      |
|------|---------------|---------------------------------------------------------|
| lv   | afrikānietis | {{suffix|lv|afrikānis|ietis|gloss11=African}} |
| la   | aequabilis   | {{af|la|aequō|alt1=aequāre, aequō|t11=I make even, level|-bilis}} |
| nl   | stokhaver    | {{compound|nl|stok|t11=stick, cane|haver|t2=oats, fodder, a feed, dose}} |
| nl   | versnelling  | {{suffix|nl|versnellen|t1=accelerate|ing|t22=ation}} |

Table 3: Template gloss argument with typos bolded.

### 4 Extracting Morphological Information

Wiktionary is also a rich source of morphological information. Here we focus on one type of information, which we call “form-of relations” because they are annotated in Wiktionary using Form-Of templates. We extract 4,027,201 relations across 168 relation types, a full histogram of which is in Appendix A. While different relations have different requirements as to where they can appear in an entry (e.g. some relations can only appear in the etymology section), form-of relations are relatively straightforward to extract and normalize due to the consistency of their templates.

Many inflectional relations for both nouns and verbs, including relations such as inflection-of, genitive-singular-of, or past-participle-of, are already packaged in UniMorph and have been used in tasks such as morphological inflection analysis and prediction (McCarty et al., 2019; Kann et al., 2020). Other relations, such as plural-of and feminine-form-of can augment training data for morphological analysis systems such as that of Nicolai and Yarowsky (2019). However, much of the rest of this form-of data has not been thoroughly explored. Below, we present preliminary experiments on clipping, contraction, and

---

1 Rendered in HTML as: from Latin *ab* (“from, away from”)
2 A comprehensive list is at [https://en.wiktionary.org/wiki/Category:Form-of_templates](https://en.wiktionary.org/wiki/Category:Form-of_templates)
eye dialect, three understudied types of data whose further research is enabled through our extraction and normalization.

4.1 Experiments

We experiment with predicting three form-of relations. **Clipping** is a process of word formation in which a part of the word gets “clipped” or truncated to form a new word that retains both original word’s meaning and part of speech. Common examples in English include *math* from *mathematics* or *phone* from *telephone*. **Contraction** occurs when sounds or letters are dropped to form a new, shorter word or word group. In English, examples include *I’m* from *I am* and the bound morpheme *-n’t* from *not*. **Eye dialect** is the use of nonstandard spelling to highlight a word’s pronunciation. It is often used in literary works to draw attention to a character’s particular dialect or accent. Some examples in English include *aftuh* for *after* and *jokin’* for *joking*. In Wiktionary, several eye dialect annotations include the specific dialect represented, such as African American Vernacular English (AAVE) or Southern US.

For these linguistic phenomena, Wiktionary contains annotations across a wide range of languages. The amount of annotations is also quite small: the total amount of data is only around 1-2K instances per task (Table 4). While there has not been much published computational literature on these tasks, we envision interesting potential downstream applications for systems successful at generating clippings, contractions, and eye dialectical variations. For example, changing the language style of chatbots has been shown to increase user satisfaction (Elsholz et al., 2019).

**Models**  We use a character neural machine translation setup. Using OpenNMT-py (Klein et al., 2017), we employ a 2-layer LSTM encoder-decoder with 256-dimension hidden and embedding size, batch size 64, Adam optimizer with learning rate 0.001, and patience of 5. We train two model variants, a baseline with Luong attention (Luong et al., 2015) (the default in OpenNMT), and a second with copy attention (Gu et al., 2016). For eye dialect, we only use English data, as the overwhelming majority of annotations are English. For clipping and contraction, we employ the entire range of languages annotated, thus making our models multi-source, multi-target systems. We use a randomly shuffled 80-10-10 train-dev-test split. The input and output format of each experiment, as well as results are presented in Table 5.

| Task          | Top 5 languages (count) | Total Languages |
|---------------|-------------------------|-----------------|
| Clipping      | en (575), ja (246), pt (118), de (67), fr (56) | 1461 57        |
| Contraction   | en (414), pt (96), de (79), dum (63), ga (50)   | 1404 82        |
| Eye Dialect   | en (1646), pt (149), vi (89), da (35), es (32)  | 2064 39        |

Table 4: Total available data for each tasks, including top five languages. Only English data was used for eye dialect experiments.

| Experiment | Input Format | Output Format | Luong Attn | Copy Attn |
|------------|--------------|---------------|------------|-----------|
|            | 1-best       | 5-best        | 1-best     | 5-best    |
| Clipping   | ht k a p a b  | k a p         | .25 (2.5)  | .29 (2.0) |
|            |              |               | .38 (2.1)  | .49 (1.5) |
| Contraction| en parents   | ’rents        | .35 (1.7)  | .49 (1.2) |
|            |              |               | .39 (1.5)  | .54 (0.9) |
| Eye Dialect| twenty       | twenty        | .32 (1.6)  | .42 (1.1) |
|            |              |               | .39 (1.5)  | .48 (1.0) |

Table 5: Experimental results. Metrics are exact match accuracy and (mean character edit distance).

**Results**  We compute exact match accuracy and average character edit distance to the gold for each setting. Though 1-best and 5-best accuracies across all three tasks seem low, actually on average the results are only 1–2 characters off from the gold; we see the model consistently making plausible predictions with similar sounds. In addition, the models with copy attention consistently outperform the models with a standard Luong attention. Due to space constraints, sample predictions are presented in Appendix B, and improvements of the copy attention model over the Luong attention model are in Appendix C.

---

3For monotonic sequence-to-sequence tasks, LSTMs tend to perform better than Transformers (Gorman et al., 2020).
Analysis  Clippings tend to keep the beginning part of the word (speculation → spec), which the model learned (Spotlight → Spot), albeit sometimes incorrectly (Alfredino → Alfe, gold is Dino). A large percentage of clippings are in Japanese; if the input is written in katakana, the model can sometimes make a correct prediction, but if written in kanji, the model gets it completely wrong, due to the rarity of the characters. These errors are corrected by the copy attention model, which learns to copy over characters that would otherwise be unlikely to be generated. Contraction is perhaps an easier form of clipping; the model learns to keep characters at the beginning and end of a word. For eye dialect, the models successfully learned the -ing → -in’ mapping. We observe that many incorrect predictions are often quite acceptable to a human depending on one’s dialect of English (old → ole, gold is owld; yourself → yoself, gold is youself). Thus character-based metrics may be more informative measures of performance than accuracy. Overall, the copy attention model substantially outperforms a regular attention baseline, due to the fact that the output contains many characters from the input (for clipping and contraction, the task is akin to selecting characters to keep and or discard).

5 Conclusion

We extend a Yawipa, a comprehensive Wiktionary parser, to extract and normalize translations from etymology glosses and morphological form-of relations, resulting in substantial increases in extracted data. Our multilingual neural sequence models trained on very low amounts of data show quite low character edit distance when predicting words formed through clipping, contraction and eye dialect. We show that copy attention works well for tasks where the output is a mutation of the input. We envision our newly extracted data to be extremely valuable to researchers working with multilingual text data. Data and code are available at github.com/wswu/yawipa.

References

Judit Ács. 2014. Pivot-based multilingual dictionary building using Wiktionary. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 1938–1942, Reykjavik, Iceland. May. European Language Resources Association (ELRA).

Julian Brooke, Tong Wang, and Graeme Hirst. 2011. Predicting word clipping with latent semantic analysis. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 1392–1396, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, pages 301–305.

Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. 2020. A spelling correction corpus for multiple Arabic dialects. In Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, May. European Language Resources Association.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 40–50, Online, July. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1631–1640, Berlin, Germany, August. Association for Computational Linguistics.

Marcelo Yuji Himoro and Antonio Pareja-Lora. 2020. Towards a spell checker for zamboanga chavacano orthography. In Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, May. European Language Resources Association.

Katharina Kann, Arya D. McCarthy, Garrett Nicolai, and Mans Hulden. 2020. The SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. In Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 51–62, Online, July. Association for Computational Linguistics.
Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 3121–3126, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May. European Language Resources Association (ELRA).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Vivek Kulkarni and William Yang Wang. 2018. Simple models for word formation in slang. 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 1424–1434, New Orleans, Louisiana, June. Association for Computational Linguistics.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 4223–4228, Marseille, France, May. European Language Resources Association.

Matthias Liebeck and Stefan Conrad. 2015. IWNLP: Inverse Wiktionary for natural language processing. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 414–418, Beijing, China, July. Association for Computational Linguistics.

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, September. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shi Jiwe, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 229–244, Florence, Italy, August. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangel'skiy, Natalya Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ermstreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal morphology. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 3922–3931, Marseille, France, May. European Language Resources Association.

Clifton McFate and Kenneth Forbus. 2011. NULEX: An open-license broad coverage lexicon. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 363–367, Portland, Oregon, USA, June. Association for Computational Linguistics.

V. Nastase and C. Strapparava. 2015. knowitiary: A machine readable incarnation of wiktionary. Int. J. Comput. Linguistics Appl., 6:61–82.

Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Ivy Kuo, Pierre Magistry, and Chu-Ren Huang. 2009. Wiktionary for natural language processing: Methodology and limitations. In Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources (People’s Web), pages 19–27, Suntec, Singapore, August. Association for Computational Linguistics.

Garrett Nicolai and David Yarowsky. 2019. Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1765–1774, Florence, Italy, July. Association for Computational Linguistics.

Franck Sajous, Basilio Calderone, and Nabil Hathout. 2020. ENGLAWI: From human- to machine-readable Wiktionary. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 3016–3026, Marseille, France, May. European Language Resources Association.
Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Martin Volk and Rico Sennrich. 2011. Disambiguation of English contractions for machine translation of TV subtitles. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 238–245, Riga, Latvia, May. Northern European Association for Language Technology (NEALT).

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France, May. European Language Resources Association.
A Form-Of Histogram

A histogram of all form-of relations we extracted from Wiktionary. This paper experimented with clipping, contraction, and eye dialect.

3026829 inflection of
437008 plural of
92330 alternative form of
49974 present participle of
38753 feminine singular of
35914 feminine plural of
31855 masculine plural of
27673 alternative spelling of
24350 past participle of
16420 synonym of
13927 gerund of
12555 definite singular of
11333 initialism of
11276 romanization of
9130 abbreviation of
8434 superlative of
8029 diminutive of
7727 comparative of
7455 female equivalent of
6970 masculine plural past participle of
6926 feminine singular past participle of
6786 feminine plural past participle of
6771 misspelling of
6499 obsolete spelling of
6006 justspringing reading of
5244 obsolete form of
5221 definite plural of
4881 indefinite plural of
4216 verbal noun of
4054 form of
3723 genitive of
3584 genitive singular of
2587 present tense of
2583 passive of
2580 ady form of
2063 eye dialect of
1986 dative plural of
1976 archaic form of
1671 nonstandard spelling of
1667 reflexive of
1621 imperative of
1617 dative of
1544 alternative case form of
1510 short for
1461 clipping of
1404 contraction of
1389 neuter singular of
1161 acronym of
954 imperfective form of
945 archaic spelling of
928 past tense of
923 eclipse of
908 soft mutation of
897 apocopated form of
799 dated form of
779 standard spelling of
774 vocative singular of
753 superseded spelling of
699 attributive form of
631 spelling of
602 rare spelling of
503 augmentative of
485 nasal mutation of
406 rare form of
379 singularative of
375 da-e-form of
368 elipsis of
364 illusion of
356 neuter singular past participle of
353 b-prothesis of
333 aspirate mutation of
329 en-archaic second-person singular past of
278 morse code for
257 participle of
238 elative of
220 agent noun of
218 nominative plural of
216 nonstandard form of
186 dated spelling of
182 pronunciation spelling of
179 negative of
158 misconstruction of
156 medieval spelling of
150 former name of
144 feminine of
133 enduring form of
130 ru-abbrev of
129 nasless form of
127 yi-unpointed form of
104 active participle of
103 dative singular of
103 causative of
102 genitive plural of
100 ru-initialism of
100 obsolete typography of
97 superlative predicative of
95 superlative attributive of
95 informal form of
76 elongated form of
73 euphemistic form of
68 passive participle of
68 alternative plural of
68 alternative typography of
68 alternative plural of
61 el-pol of
60 peorative of
54 t-prothesis of
54 perfective form of
52 singular of
50 pt-superseded-parexstone
50 euphemistic spelling of
47 uncommon spelling of
47 past active participle of
47 harmonic variant of
46 superseding form of
43 syncopic form of
41 abstract noun of
40 supine of
37 dual of
35 en-ing form of
35 eggcorn of
34 informal spelling of
33 ru-acronym of
29 equative of
28 hard mutation of
27 slender form of
25 standard form of
25 iterative of
24 accusative singular of
24 accusative plural of
23 common form of
23 future participle of
20 deliberate misspelling of
18 past passive participle of
18 honorific alternative case form of
17 mixed mutation of
16 vocative plural of
16 la-pronominal abbreviation of
15 nomen sacrum form of
15 apthic form of
11 nominalization of
10 yi-phonetic spelling of
9 perfect participle of
9 my-ICT of
9 frequentative of
8 el-mono-of
6 masculine of
6 uk-pre-reform
5 pronunciation variant of
5 present active participle of
5 fi-post-1950
5 broad form of
4 pt-pronoun-with-n
4 pt-pronoun-with-l
4 diminutive plural of
4 accusative of
3 neuter plural of
3 men's speech form of
2 misromanization of
2 masculine noun of
2 epy-alternative transliteration of
1 yi-alternatively pointed form of
1 xiaojing spelling of
1 rfform
1 sh-honorific
1 rfform
1 morse code prosign
1 morse code abbreviation
1 hy-reformed
1 ceb-superseded spelling of
1 alternative reconstruction of

B Form-Of Predictions

This section contains form-of predictions by the Luong attention model. Predictions of the copy attention model look similar and often better (i.e. closer to the gold). The input for each experimental setup is character separated (with an extra leading language token for clipping and contraction). Spaces are replaced with underscores. For comparisons between the two models, see Appendix C.
### B.1 Clipping

| Input                    | Gold       | 5-best                |
|--------------------------|------------|-----------------------|
| en romantic, comedy      | romcom     | rom_com, rom-com, romicom, romac, romacom |
| f rinstein              | insti      | insto, insti, inti, int, int |
| en homosexual            | homo       | homo, pomo, tomo, sono, nomo |
| da Sebastian             | Bastian    | Seb, beb, Beb, Ses, hes |
| de Spotlight             | Spot       | Sopo, Lopo, Loso, Hopo, Loto |
| ca pagina_web            | web        | ping, pong, peng, pig, p-ng |
| eo la irlanda lingvo     | irlanda    | éir, cranana, ér, cranana, éran, ér, cranana, éran, ér, cranana, éran, ér, cranana |
| it Alfredino             | Dino       | Alfe, Alff, Iff, Iffr, Afi |
| en speculation           | spec       | spec, spec, spec, spec, spec, spec |

### B.2 Contraction

| Input                    | Gold       | 5-best                |
|--------------------------|------------|-----------------------|
| de so, une               | sone       | sonne, sonnie, so’ne, sowne, some |
| fr celui                 | qui        | chui, ccui, chai, chui, ccui |
| en about                 | abt        | abtu, abt, abut, bout, baut |
| it dalla ara             | Dall’Ara   | dra, d’ra, dral’r, d’al’r |
| en have, some             | hasome     | have’s, ha’ve, have’m, ha’smer |
| en they, will             | they’ll    | they’ll, them’ll, thea’ll, they’ll, the’yl |
| af toe, het               | toe’t      | to’t, tho’t, toe’t, the’t |
| sw huna, jambo            | hujambo    | hajambo, handamo, hamambo, hijambo, hajamo |
| en wicketkeeper          | wickie     | wiveret, whikente, whient, whievent, whieven |
| ga faoi, an              | faoin      | faoin, fao’n, fa’an, faoin, afoin |

### B.3 Eye Dialect

| Input                    | Gold       | 5-best                |
|--------------------------|------------|-----------------------|
| off                      | offn       | hoff, oof, haff, off, huff |
| cooking                  | cookin’    | cookin’, coukin’, cookin’, sookin’, coopin’ |
| gallivanting             | gallivantin’ | gallinth, gawlintin’, gawlint, gaglintin’, gawlin’ |
| raving                   | ravin’     | ravin’, rain’, rawin’, rafin’, ravin’ |
| lynching                 | lynchin’   | lanchin’, lanchin’, lagnhin’, lanchin’, lantlin’ |
| developing               | developin’ | devlopin’, devolin’, devlosin’, devlenin’, devlopin |
| yourself                 | youself    | youself, youself, thi, sen, yo’self, dosel |
| old                      | owld       | ole, old, ol’, old, wold |
| Ms                       | Miz        | mizz, Mizz, izzo, mizz, zizz |
| your                     | yur        | yor, yer, ye, yo, yire |

### C Model Improvements

This section presents sample predictions where the Luong attention model predicted incorrectly, and the copy attention model predicted correctly, showing that copy attention is useful for tasks like ours where the input and output share common tokens.

#### C.1 Clipping

| Input                    | Gold       | Luong Attn 5-best | Copy Attn 5-best |
|--------------------------|------------|-------------------|-----------------|
| fr instituteur           | insti      | insti, insti, inti, insti, intit |
| en subdebutante          | subde      | subde, subde, subde, subde, subde |
| li geograph              | geo        | geog, meg, gerg, gerg, gerga |
| en maximum               | max        | maci, maxi, mamia, mapi, mali |
| en radical               | rad        | rada, radia, radia, radia |
| tl Corazon               | Cora       | Coro, Corona, Coro, Cordo, Coro, Coro,Coro,Coro,Coro,Coro |
| eo la, itala lingvo      | itala      | ital, itala, ita, itala, itala, itala |
| en steady                | stead      | stead, stead, stead |
| eo la japana lingvo      | japana     | japana, japana, japana, japana, japana, japana |

4691
### C.2 Contraction

| Input         | Gold  | Luong Attn 5-best        | Copy Attn 5-best       |
|---------------|-------|--------------------------|------------------------|
| de weißt du   | weiße | weíite,weiête,weidti,weit,weirt | weiıte,weiıte,weiıta,weııte,weııme |
| oc de Eths    | deths | deth,detha,dethi,deths,det’s | deths,deth,deths,deths,deth |
| ro prin_o     | prinr-o | prinr-un,prinr-o,prinr-on,pirron,prinr-in | prinr-o,prí-o,prí-r,prin-run,prína |
| ca a el       | al    | as,al,a,ae,at            | al,al’,‘ll,all,’l   |
| en overlook    | o’erlook | o’erload,o’erloan,o’erloan,o’erload,o’erlo | o’erlook,o’erloo,oarlor,oarloo,oarloo |
| en overhead    | o’erhead | o’errese,o’erread,o’erread,o’erread,o’errehead | o’errehead,oerhea’,oerhe’d,oerhea’,oerhead |
| oc per Eths   | peths | peth,petha,pethas,pech,pethe | peths,prths,prths,peth,preth |
| cy eich       | ‘ch   | echi,ee’,sech,’ch,dei    | ‘ch,c’ch,chi,‘c,chi’ |

### C.3 Eye Dialect

| Input | Gold  | Luong Attn 5-best | Copy Attn 5-best |
|-------|-------|-------------------|------------------|
| lynch | lynch’ | lanchin’,lyanchin’,lanthin’,lanychin’,lanthin’ | lynchin’,lyanchin’,lyanching,lunchin’lunchin’ |
| baptiz | baptizin’ | baptin’,baptizin’,bastin’,bastin’,baptizin’,baptizin’ | baptizin’,bawtizin’,baptizin,baptizing,baptizin’ |
| grazin | grazin’ | grazin’,grazin’,grazin’,grazin’,grazin’ | grazin’,grazin’,grazin’,grazin’,grazin’ |
| mutatin’ | mutatin’ | mutatin’,mutatin’,mutatin’,mutatin’,mutatin’,mutatin’ | mutatin’,mutatin’,mutatin’,mutatin’,mutatin’,mutatin’ |
| insultn’ | insultn’ | insultn’,insultn’,insultn’,insultn’,insultn’ | insultn’,insultn’,insultn’,insultn’,insultn’,insultn’ |
| amazin’ | amazin’ | amazin’,amazin’,amazin’,amazin’,amazin’ | amazin’,amazin’,amazin’,amazin’,amazin’,amazin’ |
| pukin’ | pukin’ | pukin’,pukin’,pukin’,pukin’,pukin’,pukin’,pukin’,pukin’ | pukin’,pukin’,pukin’,pukin’,pukin’,pukin’,pukin’,pukin’ |
| repeatin’ | repeatin’ | repeatin’,repeatin’,repeatin’,repeatin’,repeatin’,repeatin’ | repeatin’,repeatin’,repeatin’,repeatin’,repeatin’,repeatin’ |
| ‘onour | ‘onour | ‘onour,‘onour,‘onour,‘onour,‘onour,‘onour,‘onour,‘onour | ‘onour,‘onour,‘onour,‘onour,‘onour,‘onour,‘onour,‘onour |
| pumpin’ | pumpin’ | pumpin’,pumpin’,pumpin’,pumpin’,pumpin’,pumpin’,pumpin’,pumpin’ | pumpin’,pumpin’,pumpin’,pumpin’,pumpin’,pumpin’,pumpin’,pumpin’ |