Mining the Spoken Wikipedia for Speech Data and Beyond

Arne Köhn, Florian Stegen, Timo Baumann
Department of Informatics
Universität Hamburg
{koehn,1stegen,baumann}@informatik.uni-hamburg.de

Abstract

We present a corpus of time-aligned spoken data of Wikipedia articles as well as the pipeline that allows to generate such corpora for many languages. There are initiatives to create and sustain spoken Wikipedia versions in many languages and hence the data is freely available, grows over time, and can be used for automatic corpus creation. Our pipeline automatically downloads and aligns this data. The resulting German corpus currently totals 293h of audio, of which we align 71h in full sentences and another 86h of sentences with some missing words. The English corpus consists of 287h, for which we align 27h in full sentence and 157h with some missing words. Results are publically available.[1]

Keywords: Spoken Wikipedia, Long Audio Alignment, Speech Resource

1. Introduction

Most time-aligned speech data for corpus analyses or training models is non-free. This produces a barrier for research. In addition, models generated from these corpora cannot be freely distributed which also hinders research.

The Spoken Wikipedia[2], in contrast, is a large speech resource under a free license with corresponding text available, covering a broad variety of topics. It is constantly evolving and of considerable size for several languages.

While the (written) Wikipedia is already being widely used for research in computational linguistics (Atserias et al., 2008; Ahn et al., 2004; Nothman et al., 2009; Horn et al., 2014, and many others), the Spoken Wikipedia is not yet used in speech research, although it is a broad and multilingual source with lots of automatic and manual annotation available (such as links and topic relatedness) for the textual material. One major problem is the missing linkage between the spoken and written text as well as the semi-structured nature of Wikipedia data.

We present a pipeline that aligns audio from the Spoken Wikipedia to the article text being read. We show that the data and process allow to bootstrap free speech recognition models from non-free ones (which we will publish with the full paper).

The remainder of the paper is structured as follows: we present some descriptive statistics about the Spoken Wikipedia corpus in Section 2, highlighting the overall amounts of material that is available. We then describe some challenges about the available data in Section 3 and our pipeline for extracting and editing the downloadable data into a useful resource in Section 4. We describe the resulting data sets in Section 5, will describe our use of the data in Section 7 of the full paper and conclude with Section 8.

2. Spoken Wikipedia

The Spoken Wikipedia is a project in which volunteers read out and record Wikipedia articles. One main aim is to make the information from the articles available through another modality, e. g. for visually impaired people, and for other hands- and eyes-free uses, such as while driving.

The articles being read are of course not randomly selected but depend on the interests of the readers. For example, so called "stub" articles are underrepresented and noteworthy articles are overrepresented.

Using data from the Spoken Wikipedia has several advantages: The articles are read by a large and diverse set of people, cover a variety of topics such as cities (Ingolstadt, 152 minutes), famous researchers (Carl Friedrich Gauß, 54 minutes), technical articles (Microsoft Windows NT 3.1, 67 minutes), mathematics (number theory, 22 minutes), and are not only available free of charge but actually licensed under a creative commons (CC-by-SA) license. Although the Wikipedia is constantly evolving, the audio files are co-referenced with the exact article revision that was being read (with some exceptions, see below), allowing to match the audio with the read text.

Spoken data is available for 28 languages, with English, Dutch and German being the largest collections. We focus on German in this paper as there is otherwise relatively little free speech material available for German. We have also validated our generalizations by looking at English data; our software pipeline described below also works for Dutch and can be extended for other languages.

The German Spoken Wikipedia so far contains 864 spoken articles read by 299 identified speakers (speaker information is missing for some speakers), totaling 293 hours of audio (of which 35 hours are missing speaker information). A plot of speaker contributions is presented in Figure 1 (upper red points). As the figure shows, the distribution of contribution (by audio duration) is highly skewed. Very few speakers speak tremendous amounts of audio. Although many speakers only read a single article (not emphasized in the figure), most are still represented by at least 10 minutes of audio.

The English Spoken Wikipedia so far contains 1240 spoken articles read by 413 identified speakers (speaker information is missing for some speakers), totaling 287 hours of audio (indicating that the read articles a shorter on average than in the German Spoken Wikipedia).

[1] nats-www.informatik.uni-hamburg.de/SWC/

[2] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia; also contains links to other languages.

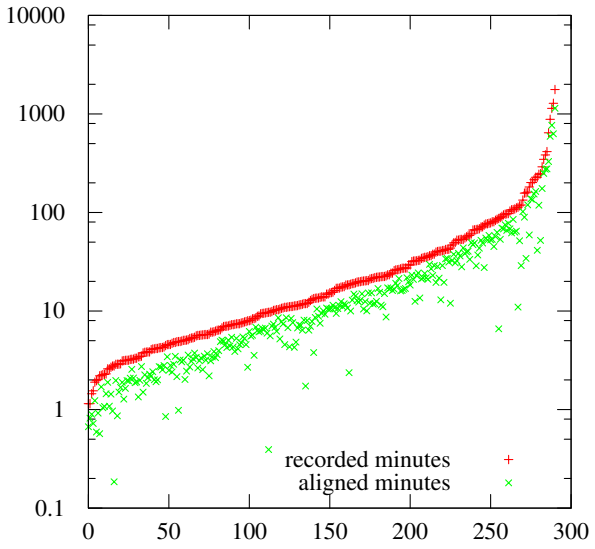

Figure 1: Contribution of audio by speaker in the German Spoken Wikipedia. Half of the speakers contribute between 10 and 100 minutes of audio; only a few contribute (much) more (up to 40 hours); except for some outliers, the alignment rate is relatively stable across speakers.

3. Challenges

There is one key challenge regarding the alignment: Although we have an audio file of the article being read as well as the article's source code, the written part does not always match what is being read. There are (1) parts in the text that are not being read, (2) parts we don't know how they are being read, and (3) there are parts in the audio that are not part of the text.

Regarding problem (1), the articles are cleaned from wiki markup. We also drop tables and boxes because they are most often not read and if they are, it is not predictable, in which manner and ordering. Most footnotes are not being read and therefore we drop them as well. This approach might err on the side of dropping too much, but it prohibits false alignments to text that is not spoken at all.

Regarding (2), there is some text where we don't know whether or how it will be spoken, e. g. the headings and mathematic formulas. We try to normalize some formulas, but the coverage is limited (see also Ferres and Sepúlveda (2011) for a solution to this problem). In the case of unsuccessful normalization, the alignment algorithm will simply not be able to find an alignment.

Regarding (3), typically, the audio starts (and often ends) with a disclaimer about the origin of the text and the license of the audio. This is not part of the article but is sufficiently similar for each audio file that we can generate the text and prepend it to the text to be aligned. Other parts will not be aligned (as the corresponding text is not available).

One of the main challenges is the constant evolution of Wikipedia: articles (or their spoken versions) are added, article text is revised, meta-information is updated or erroneously breaks for a multitude of reasons. As a consequence, our solution must not rely on manual corrections of the extracted data. Instead, our solution needed to be robust to errors, able to extend the corpus without forcing recomputations on the unchanged data, and provide systematic workarounds for missing data (until that data is added at the source).

4. Automatic Annotation Pipeline

Our pipeline works as follows: First, we scrape the Wikipedia for articles that are in the category for spoken articles. We then download the corresponding text and audio. The text needs to be normalized before it can be aligned to the audio. We will describe all these steps in this section.

4.1. Scraping the Wikipedia

All articles with a spoken version contain a template which in turn results in some info box and (for most languages) a category marker. We use the Wikipedia API to query the spoken article category, then examine the article source for the template, which contains the read article revision ID, speaker ID and reading date, as well as a link to the Wikimedia Commons page that contains the actual audio (in one or more files, most often encoded as OGG-Vorbis in high quality). The Wikipedia is a semi-structured database, which means that values can be missing or mal-formatted. Our software contains many workarounds to deal with such issues (e. g. if the revision ID is missing from the template, we estimate it from the reading date and the article history). We also started to correct missing bits in Wikipedia pages in unambiguous cases. For this, the error/warning reporting of our tool is crucial.

4.2. Text Normalization

Wikipedia pages can be downloaded in several formats including WikiMarkup and an HTML-like format (that is probably fed to the CMS). The latter is easier to convert to raw text, including stripping footnotes, "citation-needed" marks, and other Wikipedia markup. We then use MaryTTS (Schröder and Trouvain, 2003) for sentence segmentation and tokenization. Our intermediate formats ensure that the original text and and the final normalized text remain in synchrony so that timing information for the original text can later be inferred based on the alignment of the normalized text. We also add some additional text normalizations (in particular for years, some simple formulae (which are in LaTex notation) and some common units.

As a special kind of normalization, we add the spoken "header" of each article that mentions the name of the article, the license, the date it was read, and possibly by whom using a pattern filled from the meta-data. We also filter out any textual lists of references, as these are typically not read (and would be hard to normalize).

4.3. Audio Alignment

To perform the audio alignment, we employ a variant of the SailAlign algorithm (Katsamanis et al., 2011) implemented in Sphinx-4 (Walker et al., 2004) with some extensions as described below. SailAlign treats audio alignment as repeated and successively more restricted speech recognition:

The main idea is to generate an n-gram model from the text to be aligned and use this for speech recognition on the

provided audio. The speech recognition system generates a time-stamped word sequence. This sequence is matched against the original text and if a sequence of five or more words matches, the alignment for these words is kept as a landmark. The algorithm then recursively aligns the text and audio by splitting both audio and text between the landmarks and running the algorithm on each of these sub-ranges. The process stops once no new landmarks have been found. In this aspect we deviate from Katsamanis et al. (2011), who use a fixed set of iterations.

The parts that could not be aligned using the approach described above are then aligned using phone-to-phone alignment, which is more robust to errors and generates less ill-aligned sequences than forced alignment (at the cost of some coverage). We convert the text to a phone sequence using Sphinx's grapheme to phoneme conversion. We then perform a phone-based speech recognition (where the result is the most probable sequence of phones – which do not need to correspond to actual words). These two phone sequences – one generated from the text and one recognized from the audio – are then aligned using the Dijkstra algorithm (which behaves similar to the Needleman-Wunsch algorithm). The penalties for phone substitutions have been learned from aligning some of the Wikipedia data and then computing the confusion probabilities for phone pairs.

Finally, if a word remains unaligned but both preceding and succeeding words have timings, we could infer the missing timing. However, we found that such words are likely to have been mis-normalized (e. g. "*" is often spoken as "geboren" (*born*) but normalized as "Sternchen" (*asterisk*)) and should therefore not be aligned. Overall, in our implementation, we favor quality over coverage.

5. Resulting Data Sets

We aligned German and English articles and make the results available to the public.[1] The resulting aligned data is in XML format and uses a cleaned HTML rendering of the wikitext source received via the Wikipedia API as base. The text is tokenized and marked with sentence boundaries. In addition, each token is annotated with its normalized version as well as the start and end timings if it was successfully aligned.

5.1. German Data

For German, we aligned 763 articles, containing 260 hours of speech and 2.1 million tokens. We successfully aligned 157 hours of audio (60 % of the original audio) to their respective word tokens (for a total of 1.3M aligned tokens in 140k word forms). For many analyses or training procedures, fully-aligned sentences are desirable. When only looking at completely aligned sentences (i. e. every word of the sentence is aligned), we total 71 aligned hours (27 % of the original audio) in about 30k sentences.

The 288 aligned readers recorded on average 54 minutes. However, the distribution is extremely skewed, with the median at 13 minutes and even the 75%-percentile at 40 minutes (compare Figure 1, lower green points).

5.2. English Data

For English, we aligned 1066 articles, containing 287 hours of speech and 2.7 million tokens. We aligned slightly more audio than for German (184 hours), but much less in whole sentences (27 hours). We also only aligned half the number of word forms (86k). The readers recorded on average slightly less than their German counterparts (41 minutes). The distribution is again highly skewed with the 75%-percentile at only 35 minutes.

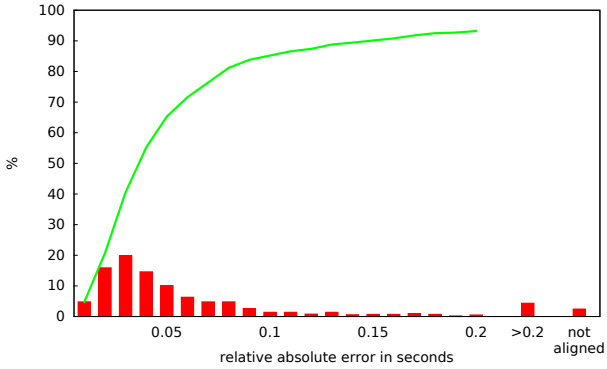

Figure 2: Distribution of absolute alignment precision compared to manual alignment. Most words that can be aligned automatically have errors smaller lower than 100 ms.

5.3. Manual Verification

We have additionally manually annotated the word-level boundaries for two spoken articles (Photodiode) containing 859 words in order to evaluate the alignment quality. In that article, the coverage is somewhat higher than average at 97.5 %. A histogram of automatic alignment deviations from the manual alignment is presented in Figure 2. As can be seen in the figure, the alignment quality is high with most words being aligned with little timing error.

6. Error Analysis

Since a relevant portion of the data could not be aligned successfully, we performed a qualitative error analysis. We found several error types which we describe in this section. Sometimes, the audio quality is simply too low for good recognition, either because of loud noise (which could potentially be filtered out) or distorting microphones. Accents such as Swiss German are also a problem for the alignment – most likely due to the acoustic models not fitting – as well as synthetic voices (which are used for some English articles). Some audio contains music, which can also not be aligned. For some articles, only a part (often the beginning of the article up to the first section) has been read at all. These problems can not be easily solved.

As noted before, the Wikipedia is semi-structured. There is no automatic consistency check for the structured data. For several articles, the wrong version ID of the article was stored. As version IDs are global (they need to survive the renaming of articles), Wikipedia then simply serves a completely different article which does not match the expected textual reference at all.[3]

Another class of errors stems from pronunciations which differ from the expected ones. The normalization fails in

[3] It should be noted though that our pipeline did not align any non-fitting text in these cases, demonstrating its robustness.

different ways (e. g. "Papst Pius XI" is spoken as "Papst Pius der Elfte", *Papst Pius the eleventh*) and loan words such as "Engagement" do not match the expected pronunciation. For words not consisting of Latin characters such as Chinese names, we don't generate a pronunciation at all.

For English, we were especially interested in the low percentage of whole sentence alignments, which is only half of the German one. It turns out that function words (e. g. *a, the, of, in,'s*) are often not aligned because they were not normalized to their reduced form.

Finally, the text normalization to just one possible pronunciation is necessarily too narrow and it might be fruitful to input multiple alternative normalization options into the alignment process. This is an area of ongoing and future work which will lead to more aligned data (and possibly into insights about text normalization in context).

7. Application to Speech Recognition

We have used a bootstrapping method in which we used some previously existing (limited quality) acoustic models for German (Baumann et al., 2010) and used these for audio alignment. We then built new models based on this data as well as the Voxforge corpus and re-aligned. While our first model was only able to align 68h, this grew with every iteration reaching 157h after 4 iterations.

Theoretically, better alignment quality after several iterations could result from overfitting the data. Although this is unlikel with so many speakers in the corpus, we also checked the alignment quality of our models on the Kiel Corpus of Read Speech (IPDS, 1994). Coverage is already very high with the first model but deviations from annotated phone boundaries in terms of RMSE continue to decrease with iterations. In other words: alignment quality continues to increase which leads to better phone-level estimates which leads to better alignment quality.

8. Conclusions and Future Work

We provide a novel time-aligned data source of considerable size based on the Spoken Wikipedia as well as a process to automatically obtain more data for a variety of languages. All data including the aligments is available under a Creative Commons license.

So far, we have used this data to train ASR models – in particular in order to bootstrap the alignment process. We are currently building an application which reads Wikipedia articles and makes use of the alignment information, e. g. to skip to the next paragraph in the audio.

We most certainly not yet maxed out the alignment coverage. For example, Tufiş et al. (2014) report much better alignment coverage on conversational data from using speaker adaptation. Correcting erroneous data in the Wikipedia and improving word normalization as well as using pronunciation alternatives should also yield more alignments.

The data that we provide could be a suitable source to create speech synthesis voices for those speakers who contributed large amounts of data (5 speakers with $>$ 2h fully aligned sentences and much more yet unaligned data available). We also plan to use the resource to explore syntax-prosody correlations which is another reason why we focus on high-quality full-sentence alignments in our work.

Acknowledgments

We would like to thank all Wikipedia authors and speakers for creating this tremendous amount of data. We also thank Felix Hennig for manual error analysis.

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., and Schlobach, S. (2004). Using wikipedia at the TREC QA track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume Special Publication 500-261. National Institute of Standards and Technology (NIST).

Atserias, J., Zaragoza, H., Ciaramita, M., and Attardi, G. (2008). Semantically annotated snapshot of the english wikipedia. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Baumann, T., Buß, O., and Schlangen, D. (2010). InproTK in Action: Open-Source Software for Building German-Speaking Incremental Spoken Dialogue Systems. In *Proceedings of ESSV*, Berlin, Germany.

Ferres, L. and Sepúlveda, J. F. (2011). Improving accessibility to mathematical formulas: The wikipedia math accessor. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, W4A '11, pages 25:1–25:9, New York, NY, USA. ACM.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463, Baltimore, Maryland, June. Association for Computational Linguistics.

IPDS, I. (1994). The Kiel corpus of read speech. CD-ROM.

Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., and Narayanan, S. (2011). Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.

Nothman, J., Murphy, T., and Curran, J. R. (2009). Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 612–620, Athens, Greece, March. Association for Computational Linguistics.

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(3):365–377, October.

Tufiş, D., Ion, R., Ştefan Dumitrescu, and Ştefănescu, D. (2014). Large smt data-sets extracted from wikipedia. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., and Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. Technical report, Mountain View, CA, USA.