METL – Project
Automatic orthographic simplification

Alexandre Kabbach
alexandre.kabbach@unige.ch

Paola Merlo
paola.merlo@unige.ch

23.04.2020

Note: as much as possible, I included pdf versions of all the references cited below in the enclosed lit *directory (except when I could not find available versions online). This also includes the* Dive into Deep Learning *book of (Zhang et al., 2020) which should quickly become your new bible.*

1 Summary

In this 4-week long group project we will explore the benefits of using `seq2seq` models for *automatic ortographic simplication*, namely, for transcribing text from traditional English spelling to `SoundSpel` (Rondthaler and Lias, 1978), a proposed simplified spelling system for English illustrated in (1). The core of the project will be dedicated to crafting the best performing `seq2seq` model for the task at hand.

(1) a. Let fall upon its back the soot that falls from chimneys
Let faul upon its bak th suut that fauls from chimnys

b. Slipped by the terrace, made a sudden leap
Slipt bi th terris, maed a suden leep

c. And seeing that it was a soft October night
And seeing that it wuz a soft Oktoeber niet

2 Motivations

Writing systems are to be distinguished from *language* as a biological device (Chomsky, 1965; Lenneberg, 1967), in that they constitute man-made tools consciously developed to transcribe speech into text.[1] In the case of alphabetical systems such as English or French, *ortography* refers to the set of conventions used in writing—such as the rules governing *spelling*—which are often motivated by considerations over straightforward phoneme-to-grapheme correspondence.

However, for various historical and/or sociological reasons, orthographic systems may evolve or freeze into *complex* states riddled with idiosyncrasies that deviate significantly from the ideal one-to-one phoneme-to-grapheme correspondence.[2][3] As a result, they turn into what the philosopher Ivan Illich has called *non-convivial tools* (Illich, 1973).

This notion of orthographic complexity has been refined in the psycholinguistics literature through the notion of *orthographic depth*, which denotes the degree of correspondence between single graphemes and phonemes in a language's orthography (Frost and Katz, 1989). Figure 1 illustrates differences of mapping between orthographic and phonological networks for both *shallow* (Serbo-Croatian) and *deep* (English) orthographic systems.

Orthographic depth has been associated with a wide range of negative cognitive effects, given that orthography shapes the perception of speech (Ziegler and Ferrand, 1998; Frith et al., 1998) and influences

[1] For a (historical) overview of writing systems, see (Coulmas, 1991; Coulmas, 1999; Fischer, 2001)

[2] A well-known linguistics joke states that the following combination of letters in English—*ghoti*—can be read as *fish*, considering *gh* as in *laugh*, *o* as in *women* and *ti* as in *nation*.

[3] For a concise introduction to the history of English spelling, see Chapter 5 in (Brown, 2019). Else, for a comprehensive overview, see (Venezky, 1999; Upward and Davidson, 2011; Scragg, 2011; Crystal, 2012)

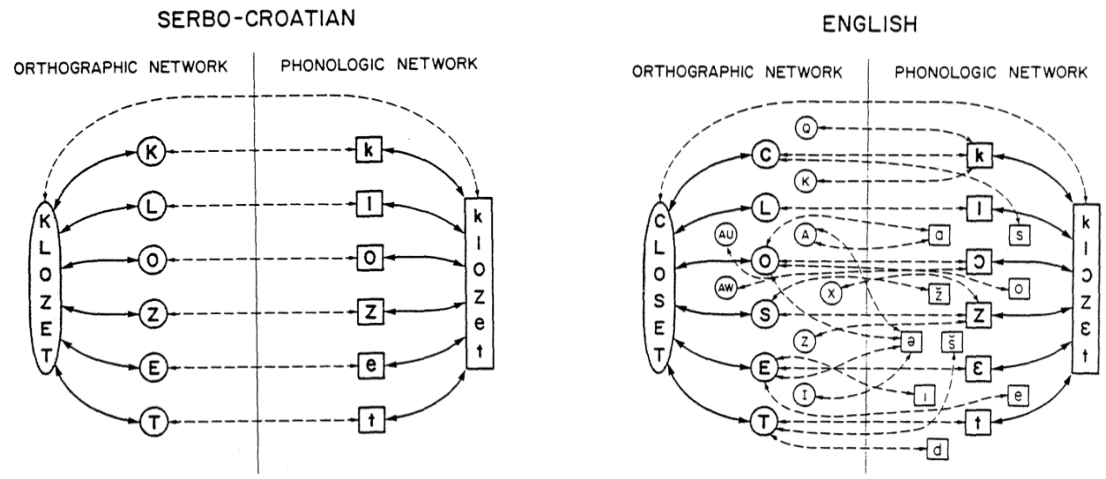

Figure 1: Interaction between orthographic and phonologic networks in English and Serbo-Croatian, from (Frost and Katz, 1989).

how literacy develops in the brain (Dehaene, 2009). More specifically, deep orthographic systems have been shown to significantly hinder the acquisition of reading in children (Oney and Goldman, 1984; Thorstad, 1991; Seymour et al., 2003), a detrimental effect on literacy that is particularly prevalent for children suffering from dyslexia (Paulesu et al., 2001; Ziegler and Goswami, 2005; Fischer et al., 2007).[4][5]

Several proposals have been made to simplify the English orthographic system.[6] In May 2019, the *English Spelling Society*[7] launched a call for proposals of alternative spelling systems for English[8] and shortlisted six of them. However, all those proposals exhibit various shortcomings. First, they represent various degrees of compromise between deep and shallow orthographic systems, given that moving from a deep to a fully shallow orthographic system has a non-negligible cost for already proficient adult readers (Dehaene, 2009). Second, they all embody a somehow top-down approach to linguistic conventions at large, and as such they also operate arbitrary decisions motivated solely by the perception of the architect of the new norm, glossing over phonetic differences across varieties of English, regionalisms, or even sociolects of different speech communities (Wardhaugh and Fuller, 2015). Last, they face some concrete technological obstacles to their widespread use, given that the tools build to smoothen up the transition to the new spelling system and accompany already proficient readers are mostly rule-based and as such time-consuming and not flexible.

In this project, we posit that recent developments in deep learning allow us to build robust models for inferring spelling simplification rules directly from text data. As such, we argue that this should allow us to follow a more bottom-up and usage-based approach to language that should prove better able at incorporating and adjusting to linguistic and normative variations across both speech and cultural communities. More specifically, we hypothesize that the problem of automatic orthographic simplification can be approximated to a problem of (character-based) neural machine translation, in that it is equivalent to converting a sequence of arbitrary length to another sequence of arbitrary and possibly distinct length, and propose to investigate the adequacy of `seq2seq` models for the task at hand.

[4] According to (Lyon et al., 2003): *Dyslexia is a specific learning disability that is neurobiological in origin. It is characterised by difficulties with accurate and/or fluent word recognition and by poor spelling and encoding abilities. Their difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction.*

[5] Curious readers may wonder about mixed syllabo and/or morpho-logographic writing systems such as Japanese or Chinese. For such systems, reading is processed differently cognitively, see (Leong and Tamaoka, 1998; Tan et al., 2003)

[6] For a brief overview see https://en.wikipedia.org/wiki/English-language_spelling_reform, else see (Crystal, 2010).

[7] http://spellingsociety.org/

[8] http://spellingsociety.org/international-english-spelling-congress

3 Tasks

3.1 Understanding the data: annotating

Your first task will be to create a test set against which to evaluate your model. This will help you familiarize yourself with the `SoundSpel` rules.

To do so, you will proceed as follows: under the `data` directory you will find the `SoundSpel` proposal detailing its rules for phoneme-to-grapheme correspondence. You will also find a training set of tokenized tab-separated word/sentence pairs and a tokenized sample of the first article of the English Wikipedia of March 2020. The available `SoundSpel` training set was generated by extracting and preprocessing the information contained in Table B1, B2, and H1 to H6 of the `SoundSpel` proposal.

Following the structure of the training set, you will create a set of tab-separated word/sentence pairs **not** already contained in the training data. Try to create a balanced subset representative of all the rules detailed in Table A to Table F of the `SoundSpel` proposal: this will help us qualitatively analyze the performance of our models and its ability to adequately generalize/infer the rules from the training data.

Balance your test set between bare word pairs and corpus-extracted sentences, following the structure of the training set. Select sentences from the Wikipedia sample. The training set contains about 1818 word pairs in total (including word pairs contained within sentence pairs) so you should try and constitute a test set of at least 10% of that size, that is, about 200 word pairs.

The most important points being that:

1. your test data should not already be found in the training data; and
2. your test data should be balanced and comprehensive with respect to the `SoundSpel` rules.

We will compare predicted and gold sequences of tokens using the Edit (Levenshtein) Distance,[9] which we may eventually modify to obtain a normalized similarity metric (Marzal and Vidal, 1993; Yujian and Bo, 2007), modulated to also incorporate transpositions, following the Damerau-Levenshtein distance.[10]

3.2 Formulating a hypothesis: modeling

Once you have familiarized yourselves with the `SoundSpel` rules, you will turn to the core of the project, which is to try and find the best possible variant of the `seq2seq` model for the task at hand.

To do so you will rely on an adapted version of the `seq2seq` code of the `pytorch` tutorial,[11] specifically designed for our task and available here: `https://github.com/akb89/ortografix`.

The aforementioned model is a character-based model that processes sequences of word pairs. That is, it does not take into account the sentential context within which word may occur in order to convert one sequence of letters to the other.

The model follows that of the `pytorch` tutorial and contains an Encoder (see Figure 2) a standard Decoder *without* Attention (see Figure 3), and a more complex Decoder *with* Attention (see Figure 4).

We will experiment with the following variants:

1. RNN, GRU or LSTM in the Encoder and Decoder
2. with and without Attention in the Decoder
3. with and without bidirectionality in the Encoder
4. with and without stacking in the Encoder and Decoder
5. with and without Dropout in the Encoder and Decoder

[9] `https://en.wikipedia.org/wiki/Levenshtein_distance`
[10] `https://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance`
[11] `https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html`

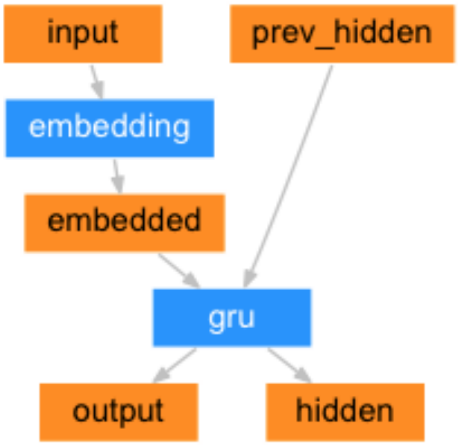

Figure 2: GRU-based Encoder

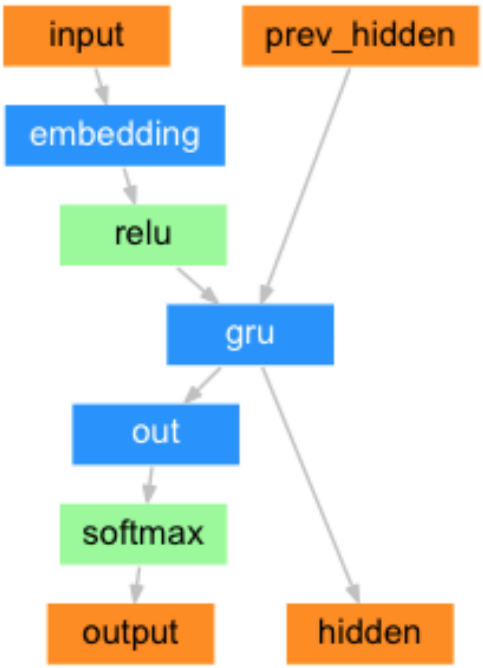

Figure 3: Simple GRU-based Decoder

Background information regarding all the aforementioned concepts is provided in the enclosed *Dive into Deep Learning* book of (Zhang et al., 2020), which provides a comprehensive overview of deep learning concepts close to the formalism we have used so far in the TP. More specifically: standard Recurrent Neural Networks (RNN) are introduced in Section 8, and more precisely in 8.3 and 8.4. Gated Recurrent Units (GRU) are introduced in Section 9.1 and Long Short Term Memory (LSTM) in Section 9.2. Stacking for deep RNN is introduced in Section 9.3, and bidirectional RNN are introduced in Section 9.4. The Encoder-Decoder Architecture and the Sequence to Sequence model are introduced in Sections 9.6 and 9.7, while Attention Mechanisms are introduced in Section 10, and for what interests us specifically in Section 10.2. Else, dropout is explained in Section 4.6.

3.3 Validating the hypothesis: experimenting

Having familiarized yourself with the problem at hand and all the aforementioned deep learning concepts, you should be capable of formulating a hypothesis regarding which architecture should perform best. Now comes the time to validate or invalidate that hypothesis by experimenting various architectures all with various hyperparameters.

You are provided with a training dataset which should take about 10min to train on a personal computer (CPU only), for about 10 epochs, leaving you ample time to test a wide range or architectures and models.

4 Output, evaluation and deadlines

You are allowed to work on this project individually or in groups (of unlimited size). However, all members of the same group shall receive the same grade. All tasks are designed to be completed within a week. Deadlines are as follows:

1. **Create the test set.** Deadline: **May 1.** The test set should take the form of a tab-separated text file uploaded on Moodle and containing *at least* 200 word pairs in the form of bare or sentence-contextualized word pairs following the data provided to you in the training set. Your test set should cover all the rules detailed in Table B1, B2, and H1 to H6 of the `SoundSpel` proposal. It should be balanced with respect to the rules covered and should not contain data already present in the

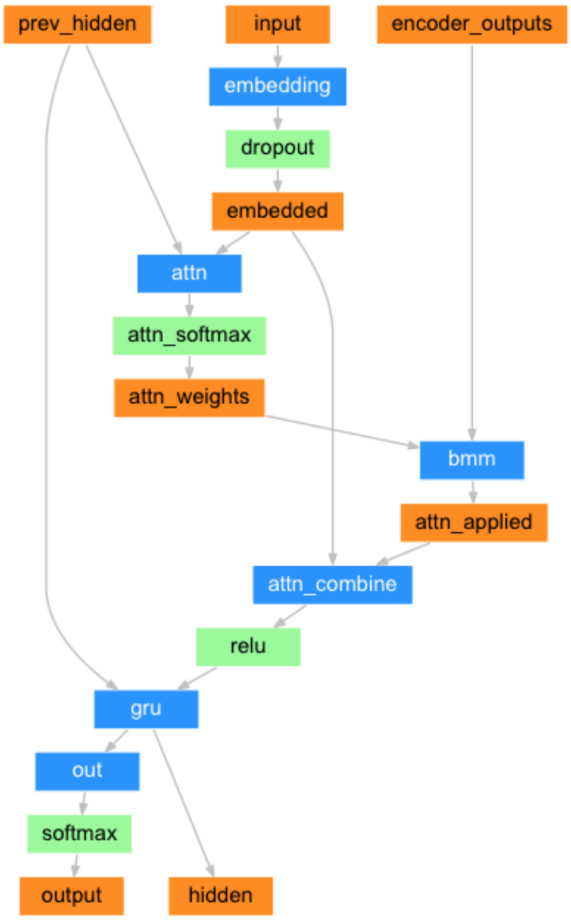

Figure 4: GRU-based Decoder *with* Attention

training set. Do not hesitate to mutualize the word between groups in order to obtain a larger test set overall. A test set of 400 word pairs will allow us to divide further between a dev and a test set and obtain a more robust experimental setup when fine-tuning the hyperparameters. Be prepared to be wrong.

2. **Formalize the hypothesis and list all possible experiments.** Deadline: **May 8.** In an up to 1 page report uploaded on Moodle, detail which variant of the model you expect to perform best on the task at hand and why. Also detail the set of hyperparameters you plan to experiment with to validate / invalidate your hypothesis.

3. **Carry out experiments and gather results.** Deadline **May 15.** Format your results into a set of LaTeXtables and upload a single PDF document on Moodle. Pay attention to the readability of your tables and the clarity and emphasis put on the important results: it should be clear from your Tables which experiments you carried out, as well as which model and which variant of each model performed best.

4. **Draft the final report.** Deadline: **May 22**. The final output of the project will take the form of a 2 to 4 pages PDF file compiled in LaTeX and uploaded on Moodle. It must follow the template provided to you on Overleaf at `https://www.overleaf.com/5825251255nynbsqkskjcp`.

We shall meet all together on zoom between 12:15pm and 1:45pm on April 23, April 30, May 7 and May 14 so that I can answer your questions and assess your progress.

References

Adam Brown. 2019. *Understanding and Teaching English Spelling: A Strategic Guide*. Routledge, New York.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Florian Coulmas. 1991. *The Writing Systems of the World*. Blackwell, Oxford.

Florian Coulmas. 1999. *The Blackwell Encyclopedia of Writing Systems*. Blackwell Publishing, Oxford.

David Crystal. 2010. *Evolving English : one language, many voices : an illustrated history of the English language*. British Library, London.

David Crystal. 2012. *Spell it out: The singular story of English spelling*. Profile Books, London.

Stanislas Dehaene. 2009. *Reading in the Brain*. Viking.

Kurt W. Fischer, Jane Holmes Bernstein, and Mary Helen Immordino-Yang, editors. 2007. *Mind, Brain, and Education in Reading Disorders*. Cambridge University Press, Cambridge, UK.

Steven Roger Fischer. 2001. *A History of Writing*. Reaktion Books, London.

Uta Frith, Heinz Wimmer, and Karin Landerl. 1998. Differences in Phonological Recoding in German- and English-Speaking Children. *Scientific Studies of Reading*, 2(1):31–54, jan.

Ram Frost and Leonard Katz. 1989. Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory & Cognition*, 17(3):302–310.

Ivan Illich. 1973. *Tools for Conviviality*. Harper & Row.

Eric H. Lenneberg. 1967. *Biological foundations of language*. John Wiley and Sons, New York.

Che Kan Leong and Katsuo Tamaoka, editors. 1998. *Cognitive Processing of the Chinese and the Japanese Languages*. Springer-Science+Business Media, Dordrecht.

G. Reid Lyon, Sally E. Shaywitz, and Bennett A. Shaywitz. 2003. A definition of dyslexia. *Annals of Dyslexia*, 53(1):1–14.

Andrés Marzal and Enrique Vidal. 1993. Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):926–932.

Banu Oney and Susan R. Goldman. 1984. Decoding and Comprehension Skills in Turkish and English: Effects of the Regularity of Grapheme-Phoneme Correspondences. *Journal of Educational Psychology*, 76(4):557–568.

E. Paulesu, J.-F. Démonet, F. Fazio, E. McCrory, V. Chanoine, N. Brunswick, S. F. Cappa, G. Cossu, M. Habib, C. D. Frith, and U. Frith. 2001. Dyslexia: Cultural diversity and biological unity. *Science*, 291(5511):2165–2167.

Edward Rondthaler and Edward J. Lias. 1978. Soundspel: A revised orthography of the English language. *IEEE Transactions on Professional Communication*, (1):25–29.

Don G. Scragg. 2011. *A history of English spelling*. Manchester University Press, Manchester.

Philip H. K. Seymour, Mikko Aro, Jane M. Erskine, and collaboration with COST Action A8 Network. 2003. Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94(2):143–174, may.

Li Hai Tan, John A Spinks, Ching-Mei Feng, Wai Ting Siok, Charles A Perfetti, Jinhu Xiong, Peter T Fox, and Jia-Hong Gao. 2003. Neural systems of second language reading are shaped by native language. *Human Brain Mapping*, 18(3):158–166, mar.

Gwenllian Thorstad. 1991. The effect of orthography on the acquisition of literacy skills. *British Journal of Psychology*, 82(4):527–537, nov.

Christopher Upward and George Davidson. 2011. *The history of English spelling*. Wiley-Blackwell, Chichester & Malden.

Richard L. Venezky. 1999. *The American way of spelling: The structure and origins of American English orthography*. The Guildford Press, New York.

Ronald Wardhaugh and Janet M. Fuller. 2015. *An Introduction to Sociolinguistics*. Wiley Blackwell, seventh edition.

Li Yujian and Liu Bo. 2007. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 2020. *Dive into Deep Learning*. https://d2l.ai.

Johannes C. Ziegler and Ludovic Ferrand. 1998. Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, 5(4):683–689.

Johannes C. Ziegler and Usha Goswami. 2005. Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, 131(1):3–29.