Building a Dictionary-Based Lemmatizer for Old Irish

Oksana Dereza
School of Linguistics, National Research University
«Higher School of Economics», Moscow, Russia
oksana.dereza@gmail.com

ABSTRACT

This paper explores the problem of developing NLP tools for morphologically rich and orthographically inconsistent classical languages. It is a case study of building a lemmatizer for Old Irish using only a dictionary and an unlabeled corpus as sources of data. At the current stage, the lemmatizer shows 76.31% average recall score on a corpus of ca. 100,000 tokens and is able to predict lemmas for out-of-vocabulary words. However, as it is the work in progress, the lemmatizer lacks some functionality such as disambiguation. There is no gold standard to measure accuracy yet either.

RÉSUMÉ

Le développement d'un programme de lemmatisation pour le vieil irlandais

Cet article vise à présenter le développement d'un logiciel de traitement automatique des langues anciennes qui sont caractérisées par une morphologie riche et par une orthographie irrégulière. Dans ce cas, il s'agit d'un outil de lemmatisation des textes en vieil irlandais créé uniquement à partir du dictionnaire et du corpus de textes non annotés. A ce stade, le rappel moyen du programme de lemmatisation est 76.31% sur un corpus d'environ 100,000 de jetons. Le programme peut prédire les lemmes pour des mots qui n'apparaissent pas dans son vocabulaire. Néanmoins, comme c'est un travail en cours, il manque encore de certaines fonctions comme la désambiguïsation sémantique. En plus, il est encore impossible de mesurer l'exactitide, parce qu'il n'y a pas de corpus annoté qui puisse servir de référence.

MOTS-CLÉS : lemmatisation, lemme, vieil irlandais, moyen irlandais, distance de Damerau–Levenshtein, données non annotées, analyse automatique de la morphologie.

KEYWORDS: lemmatisation, lemma, Old Irish, Middle Irish, Damerau-Levenshtein distance, unlabelled data, automatic morphological analysis.

1 Introduction

The interest to automatic morphological analysis of classical languages arose at the very start of computational linguistics, but still this field is underrepresented in comparison to other NLP tasks. The majority of the related works are quite old and cover only the most popular classical languages, such as Latin (Marinone, 1990), (Passarotti, 2004), ancient Greek (Packard, 1973) and Sanskrit (Verboom, 1988), (Huet, 2003). Most of the functionality of modern NLP tools for classical languages, such as CLTK,[1] is also confined to Latin and ancient Greek. However, there are many other well-documented

[1] http://docs.cltk.org/en/latest/

classical languages where a statistical-based approach to linguistic analysis may prove useful. Such languages are usually morphologically rich and orthographically inconsistent, which complicates automatic processing and requires NLP instruments to be language-specific; the lack of annotated corpora is an even bigger problem. This paper is a case study of building a lemmatizer for Old Irish using only a dictionary and an unlabeled corpus as sources of data.

2 Approach and Data

In Celtic languages, there are two ways to encode morphological information in a word form, which often occur together: initial mutations that come in the beginning of a word, and flections that come in the end. Moreover, in Old Irish some words can be incorporated into a verb between the preverb and the root: cf. *caraid* 'he / she / it loves' and *rob-car-si* 'she has loved you', where *ro-* is a perfective particle, *-b-* is an infixed pronoun for 2nd person plural object, and *-si* is an emphatic suffixed pronoun 3rd person singular feminine. The presence of a preverb with dependent forms triggers a shift in stress, which causes complex morphophonological changes and often produces a number of very differently looking forms in a verbal paradigm, particularly in the case of compound verbs, cf. *do-beir* 'gives, brings' and *ní tab(a)ir* 'does not give, bring'. This morphophonological complexity compounded by the many non-transparent features of Old Irish orthography makes the traditional dictionary approach to lemmatization with hard-coded lists of possible pseudo-suffixes and rules of their treatment less suitable for Old Irish than for other languages. A more reliable way for a start is building a full form dictionary where every word form corresponds to a lemma; an electronic edition of the Dictionary of the Irish Language[2] proves an excellent source of data for this purpose. DIL is a comprehensive historical dictionary of Irish, which covers Old and Middle Irish periods. The latest version of its electronic edition is organized in the following way: each of the 43,345 webpages is a single entry, which contains a headword, a list of possible forms and a 'main body' with translations and examples of use.

Given the aforementioned features of Old Irish, the task of building a dictionary for a lemmatizer reduces to parsing the DIL and extracting all possible forms for each lemma. However, it is not as simple as it seems. First, the list of forms cited in DIL is incomplete; it covers only about 36% of unique words in the working corpus. Second, some of the forms in DIL are contracted; for example, the list of forms for *carpat* 'chariot' looks like *cairpthiu, -thib, -tiu, -tib*. Words can be abbreviated in many different ways, which is a consequence of the fact that there were many scholars who contributed to the DIL throughout 1913-1976, and each of them used his own notation, as preserved in the digital edition.[3] Thus, one either has to drop contracted forms altogether or derive a number of rules to restore them. Third, the markup and punctuation are also inconsistent, which causes various technical problems.

The working corpus of ca. 100,000 tokens[4] was compiled from Ulster cycle sagas published on UCC CELT website.[5] It includes 24 thematically related pieces of narrative that differ in length and orthography. In the future, the corpus will be extended with texts of other forms and genres for better

[2] http://dil.ie

[3] See (Toner *et al.*, 2007) and http://dil.ie/about for details.

[4] I used a pre-trained Punkt tokenizer for English provided in NLTK, a Python library for natural language processing, which is the easiest, but obviously not the best solution for Old Irish. Building an Old Irish tokenizer is a separate important task to be solved in the future research.

[5] www.ucc.ie/celt/publishd.html

representation.

3 Algorithm and Implementation

The current version of the program consists of a dictionary compiler, a lemmatizer, and a lemma predictor for out-of-vocabulary words. The source code in Python 3 is available on GitHub.[6]

3.1 Dictionary Compiler

The dictionary compiler is a separate script that parses the DIL, extracts the list of forms for each lemma, restores contracted forms and builds a 'form : lemma' dictionary which is then dumped in JSON format for future use. It copes well with various contracted, syncopated and bracketed forms, e.g. *carat(r)as* for *caratas* and *caratras*; *carthain, -ana* for *carthana*; *cairpthiu, -tib* for *cairptib*; *caibidil, -lech* for *caibidlech*. However, the end user does not have to compile a dictionary from scratch, as its latest version always goes together with the lemmatizer.

3.2 Lemmatizer

The lemmatizer takes a file in plain text as input, cleans out punctuation and other non-word characters, and then analyses words one by one. Every word is first demutated (i.e. the changes at the beginning of the word are eliminated) and then looked up in the dictionary. The lemmatizer returns a lemma for each known word and a demutated form for each unknown word. In addition to that, the unlemmatized forms are stored in a special list. There is no word sense disambiguation for the moment, which means that if two or more different lemmas have identical forms, we cannot say for sure which lemma should be chosen for a particular instance of a homonymous form. There are two options for such cases in the current version of the lemmatizer: either return a list of all possible lemmas or choose the lemma with the highest probability. Lemma probability here equals the sum of probabilities of forms belonging to a lemma, and word form probability is a frequency count computed for each word in the corpus.

The lemmatizer has several methods, the major ones being the following:

- lemmatize a text;
- show unlemmatized words;
- evaluate performance;
- update a dictionary with a preformatted file containing new lemmas and forms.

3.3 Lemma predictor

The last module predicts lemmas for unknown words with the help of Damerau-Levenshtein distance. For every unknown word, the program generates all possible strings on edit distance 1 and 2, checks

[6] https://github.com/ancatmara/old_irish_lemmatizer

them up in the dictionary and adds those that prove to be real words to the candidate list. Then the candidates are filtered by the first character: if the unknown word starts with a vowel, the candidate should also start with a vowel, and if the unknown word starts with a consonant, the candidate should start with the same consonant. Those parameters were chosen empirically as they yield the best results, i.e. the highest percentage of correctly predicted lemmas. Finally, the lemma of the candidate that has the highest probability is taken as a lemma for the unknown word. Although this algorithm gives very promising results, it is not a default option for out-of-vocabulary words in the lemmatizer yet. There are two major reasons for this: first, the dictionary is still rather small (there are 26,160 unique tokens in the working corpus and 16,742 of them are non-dictionary forms), and second, there is no gold standard to evaluate accuracy. At the current stage, the triplets of unknown words, best candidates and their lemmas are written into an output file that requires manual revision, after which it can be uploaded as an update to the default dictionary.

The rule-based approach to lemma prediction was chosen over machine learning due to the scarcity of available data. For the moment, there are only 79,140 different forms in the 'form : lemma' dictionary compiled from the DIL, and ca. 100,000 tokens in the unlabeled Ulster cycle corpus, which is not enough for training a classifier that would be able to predict tens of different lemma classes.

4 Evaluation

As long as out-of-vocabulary words are left unlemmatized and homonymy is not taken into account, recall seems to be the most important metric for evaluating the lemmatizer's performance as it indicates the percent of forms that the program is able to process. When the recall score exceeds at least the 85-90% threshold, it will be reasonable to make a gold standard and to switch to accuracy, which is more suitable for evaluating disambiguation and unknown word treatment algorithms, because it shows the ratio of correctly predicted instances to the total number of instances.

I conducted three minor experiments to evaluate the lemmatizer's performance. First, I ran it on the whole working corpus with a default dictionary that consisted only of forms and lemmas retrieved from the DIL. This gave the average recall score of 74.7%, with the worst result of 62.5% for *Síaburcharpat Con Culainn* and the best result of 84.8% for *De Gabáil in t-Sída*.

Then I chose three random texts of different length (1,930 tokens in total, where 1,051 are unique), ran the lemmatizer with the default dictionary, manually analysed proposed lemmas for unknown words and added correctly guessed ones to the dictionary. The lemma predictor found candidates for 269 of 368 unique unlemmatized words, and 163 of them, or 61%, were correct. After that, I re-ran the lemmatizer with an updated dictionary, and the average recall score increased by ca. 10%. The results of the experiment are shown in Table 1.

Text	Tokens	Recall before update	Recall after update
Aided Óenfir Aife	1,093	79.69%	89.75%
Aided Conrói maic Dáiri	738	78.35%	89.04%
Compert Conchobuir	99	78.79%	87.88%
Overall	**1,930**	**78.94%**	**88.89%**

Table 1: Updating the dictionary with predicted lemmas

Finally, I ran the lemmatizer on the whole corpus (99,717 tokens, 26,160 unique) again, but with an updated dictionary. Although I added only 163 forms derived from only 3 texts, the recall score increased for almost every text in the corpus, the average now being 76.3%, which is 1.6% higher than before. The results also show that the score does not correlate with a text's length, but depends on the period when it was created. It is not surprising that later texts are lemmatized worse than texts written in more or less classical orthography, because the DIL contains a lot more Old Irish forms and spellings than Middle and Early Modern Irish ones. The overall results are given in Table 2.

Text	Tokens	Recall before update	Recall after update
Aided Óenfir Aife	1,093	79.69%	89.75%
Aided Conrói maic Dáiri	738	78.35%	89.04%
Aislinge Óenguso	1,267	78.61%	79.64%
Compert Conchobuir	99	78.79%	87.88%
Compert Con Culainn	1,048	83.30%	84.73%
De Chopur in dá Muccida	868	72.58%	73.27%
Do Faillsigud Tána Bó Cúailnge	326	78.53%	78.53%
Fled Bricrenn	9,006	65.33%	65.75%
Do Fogluim Chonculainn	5,486	65.33%	65.35%
De Gabáil in t-Sída	231	84.85%	84.85%
Immacallam in Dá Thúarad	637	80.69%	80.69%
Fochond loingse Fergusa meic Roig	314	75.48%	75.48%
Longes mac n-Uislenn	2,352	67.01%	67.09%
Scéla Mucce Meic Dathó	2,716	75.22%	75.85%
Mesca Ulad	7,678	76.93%	77.53%
Noínden Ulad	152	65.13%	65.13%
Serglige Con Culainn	5,943	80.67%	81.24%
Síaburcharpat Con Culainn	1,505	62.52%	62.72%
Táin Bó Fráich	3,623	80.13%	81.07%
Talland Etair	2,817	79.91%	80.90%
Táin Bó Cúailnge (Recension I)	35,744	78.78%	79.51%
Tochmarc Emire	9,576	64.00%	64.55%
Tochmarch Ferbe	6,424	71.16%	77.76%
Togail tSitha Truim	84	63.10%	63.10%
Overall	**99,717**	**74.67%**	**76.31%**

Table 2: Updating the dictionary with predicted lemmas: the whole corpus

5 Conclusion

As this is a work in progress, there are still many tasks to tackle and problems to solve. At the last run, the lemmatizer predicted lemmas for 12,156 unknown words, which is too many to filter manually. Therefore, the first priority is developing a dictionary updater that would be able to extend the dictionary with little or no human supervision. The next important task is to compile a gold standard to be able to measure accuracy and evaluate unknown word treatment and disambiguation.

The third biggest problem is disambiguation itself, which most probably requires a statistical approach. All in all, the lemmatizer is ready-to-use and shows promising results even at the current stage.

References

HUET G. (2003). Towards computational processing of sanskrit. In *International Conference on Natural Language Processing (ICON)*.

MARINONE N. (1990). A project for latin lexicography: 1. automatic lemmatization and word-list. *Computers and the Humanities*, **24**(5-6), 417–420.

PACKARD D. W. (1973). Computer-assisted morphological analysis of ancient greek.

PASSAROTTI M. C. (2004). Development and perspectives of the latin morphological analyser lemlat. *Linguistica computazionale*, **20**(A), 397–414.

TONER G., FOMIN M., BONDARENKO G. & TORMA T. (2007). : Royal Irish Academy.

VERBOOM A. (1988). Towards a sanskrit wordparser. *Literary and Linguistic Computing*, **3**(1), 40–44.