Towards Semantic Tagging for Irish

Tim Czerniak, Elaine Uí Dhonnchadha
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland
czerniat@tcd.ie, uidhonne@tcd.ie

Abstract

Well annotated corpora have been shown to have great value, both in linguistic and non-linguistic research, and in supporting machine-learning and many other non-research activities including language teaching. For minority languages, annotated corpora can help in understanding language usage norms among native and non-native speakers, providing valuable information both for lexicography and for teaching, and helping to combat the decline of speaker numbers. At the same time, minority languages suffer from having fewer available language resources than majority languages, and far less-developed annotation tooling. To date there is very little work in semantic annotation for Irish. In this paper we report on progress to date in the building of a standard tool-set for semantic annotation of Irish, including a novel method for evaluation of semantic annotation. A small corpus of Irish language data has been manually annotated with semantic tags, and manually checked. A semantic type tagging framework has then been developed using existing technologies, and using a semantic lexicon that has been built from a variety of sources. Semantic disambiguation methods have been added with a view to increasing accuracy. That framework has then been tested using the manually tagged corpus, resulting in over 90% lexical coverage and almost 80% tag accuracy. Development is ongoing as part of a larger corpus development project, and plans include expansion of the manually tagged corpus, expansion of the lexicon, and exploration of further disambiguation methods. As the first semantic tagger for Irish, to our knowledge, it is hoped that this research will form a sound basis for semantic annotation of Irish corpora in to the future.

Keywords: Irish semantic tagger, semantic annotation evaluation, Irish language NLP

1. Introduction

Semantic annotation provides a machine-readable way of understanding the meaning contained within a text, and this can be used to perform tasks such as document classification, data mining, sentiment analysis, and many other activities. According to Hovy (2022), *"While for computational linguistics, annotation is primarily an activity of corpus creation to support machine learning, for linguistics, political science, and biomedicine it can equally be a method of theory development and empirical investigation."* Annotated corpora can also be extremely useful for teaching the language in question, by providing an understanding of the norms of how a language is used in practice by its native (L1) speakers. For minority languages such as Irish, this can help maintain language use, or even increase it. Corpora also play a large part in lexicography, helping to document the current usage of the lexicon. There are existing methods for assigning part-of-speech tags to Irish texts, but while there exist several corpora of Irish containing POS annotation, there are none to our knowledge that contain semantic tags.

Irish is an Indo-European language and a member of the Goidelic branch of the Celtic language family. It has VSO (verb-subject-object) word order, fusional morphology and nominative-accusative style case-marking. Although it is the first official language of the Republic of Ireland, it is a minority language nonetheless. It has about 1.7M total speakers, of which about 77,000 in the Republic of Ireland are known to be daily speakers (CSO, 2011). It is hard to obtain a reliable number for native speakers, but reports range somewhere between 40,000 and 80,000. The language has enjoyed a recent resurgence, but it is still in a precarious position, considering the dominance of English.

Since there are few existing semantically-annotated resources for Irish, there is therefore little data to be used for machine learning (ML). The objectives of the project outlined in this paper are to start the process of building a large enough body of semantically-annotated data that could then be used for further research and ML into the future.

The main contributions of the research are a) an initial semantic lexicon for Irish, b) a small semantically tagged and checked corpus for Irish as well as c) a pipeline for semantic tagging and word-sense disambiguation for Irish texts.

The paper is structured as follows. In section 2 we explore related research, and in section 3 we describe methodology and data, including tag set selection and application to Irish. Section 4 presents a new method for evaluating semantic annotation, and discusses test results. Section 5 presents the results, section 6 explores future work and section 7 concludes.

2. Related work

Pustejovsky and Stubbs (2012) define two forms of semantic annotation: semantic typing and semantic role labelling. Each will be discussed in turn.

2.1 Semantic typing

Semantic typing involves classification of each word or phrase (i.e. lexeme) in a text. The most widely used strategies involve a hierarchical category system. To take an example, the lexeme *car* might be classified within the broad category of MOVEMENT, more specifically TRANSPORT, and more specifically VEHICLE, then LAND_VEHICLE, etc. However, there are many approaches to lexical semantics i.e. describing the meaning of words. One approach is to encode meaning via binary lexeme attributes, also called semantic features (Palmer, 1981). Taking the example of *car*, this might have the attributes +VEHICLE and +MOVEABLE, but it might also have `-ANIMAL` and `-HUMAN`, since it is neither animal nor human. A similar system is used by Bick (2000) for enhanced syntactic parsing. Some systems can also encode lexical relationships (synonymy, metonymy, hyponymy, etc.). The most widely used of these is WordNet, conceived in Miller (1995), extensively documented in Fellbaum (1998), and subsequently replicated for many languages. WordNet encodes relationships between lexemes in a very fine-grained manner, such that the hierarchical categories are part of the network itself, instead of being 'buckets' into which lexemes are grouped. Following our car example, the lexemes *car* and *vehicle* have a hyponymic relationship in WordNet. Koeva et al (2018) endeavoured to further enrich WordNet data by merging WordNet concepts with CPA (corpus pattern analysis) semantic types developed by Hanks (2004; 2013). Scannell and O'Regan (2017) document the building of a semantic network for Irish, *Líonra Séimeantach na Gaeilge* (LSG), which is modelled on WordNet. LSG is extensive in its coverage of the Irish lexicon, but much work remains to be done in verifying and checking the data therein.

2.1.1 The USAS system

The UCREL Semantic Analysis System (USAS), documented in Archer et al. (2002) and Rayson et al. (2004) was developed at the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University. The system[1] includes a set of 233 semantic categories, organised hierarchically. Each category has a top-level discourse field (denoted by an uppercase letter), a field subdivision (denoted by a number), and zero or more finer subdivisions (denoted by additional dot-separated numbers). For example, the category `N3.2` (Measurement: Size) is part of the top-level field `N` (Numbers & Measurement), and the subdivision `N3` (Measurement). A tag consists of a category, optionally followed by one or more extra symbols (`%`, `@`, `f`, `m`, `n`, `c`, `i`, `+`, `-`), which indicate semantic properties such as gender, positive or negative scale, etc. A tag may also be composed of multiple tags separated by forward-slashes, hereafter referred to as *compound tags*, indicating membership of multiple categories. As an example, the English word *chef* could be tagged with `F1/S2mf`, with double membership of the categories `F1` (Food), and `S2` (People), and because a chef can be male or female, the second tag has the extra symbols `m` and `f`. When tagging a document, a given token can be assigned a space-separated list of possible tags (each of which could be a compound tag). This assignment of multiple possible tags indicates unresolved semantic ambiguity.

The USAS system was originally developed for English, but has subsequently been extended for many languages, including one other Celtic language, Welsh as documented in Piao et al. (2015; 2016; 2017). An open-source Python library *PyMUSAS* was developed, which allows creation of lexicons for any language, and provides a tagger component that integrates with the widely used natural language processing (NLP) framework *SpaCy*[2] (Honnibal et al., 2020).

2.2 Semantic role labelling

Semantic roles describe the roles of noun phrases relative to the verb within a clause. They have been called *deep semantic cases* (Fillmore, 1968), *semantic roles* (Givón, 1990), *thematic relations* (Jackendoff, 1972) and *thematic roles* (Jackendoff, 1990). Semantic roles are critical to understanding the meaning of a clause. Each verb can be said to have a number of arguments, some required, some optional, and each dictating a particular type of noun phrase. For example the verb *hit* can be said to take a subject which has the semantic role `AGENT` (i.e. has volition), a direct object which is a `PATIENT` (i.e. it takes the effect of the verb) and can optionally have an `INSTRUMENT` as part of a prepositional phrase using *with*. Thus, *'John hit the ball with the bat'* has an `AGENT` (John), a `PATIENT` (the ball) and an `INSTRUMENT` (the bat). This 'schema' for how to use the verb *hit* is also known as a *frame*. Annotation of semantic roles generally requires a lexicon of frames for a given language, and many attempts have been made to compile these for English. FrameNet (Fillmore et al., 2003), VerbNet (Kipper Schuler, 2005) and PropBank (Palmer et al., 2005) have all produced versions of this lexicon with varying approaches. Semantic role labelling itself was formalised by Gildea and Jurafsky (2002), and this approach uses the FrameNet database. For Irish, there is as-of-yet no standard way of performing semantic role labelling, but there have been attempts to compile a frame lexicon. Wigger (2008) compiled a valency dictionary of Irish verbs[3], which could be used as a starting point for a frame lexicon to support semantic role labelling.

2.3 Semantic processing methods

As words can have many senses, one of the main issues in semantic annotation is deciding which sense is intended in a particular context. Current methods of semantic processing using neural networks and distributionally derived semantic representations (e.g. Melamud et al, 2016) while successful, require large amounts of training data which is generally not available to the majority of languages. In addition recent research has shown that hybrid models which combine both knowledge-based and distributional methods, i.e. supervised and unsupervised methods, have the most successful outcomes (Markchom et al 2023; Li & Srikumar, 2019). Therefore we concentrate our efforts initially in developing knowledge-based methods.

3. Building the semantic tagger

This section outlines the steps taken to build a semantic annotation system for Irish using PyMUSAS, SpaCy and the USAS category system,

[1] https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

[2] https://spacy.io/

[3] www.potafocal.com/fbg/

augmenting the existing Irish NLP pipeline developed in Uí Dhonnchadha and Van Genabith (2006) and Uí Dhonnchadha (2009).

3.1 Overview of the design

While semantic typing and semantic role labelling (SRL) are distinct and can be discussed separately, they are not independent. For example, classification of a noun as HUMAN would imply that it can perform the role of an AGENT. It can therefore generally be said that SRL depends on accurate semantic type annotation, and so it is wise to begin any semantic annotation framework by tackling semantic typing.

To build a semantic type tagger, we need to automatically assign a semantic category tag to each token of any given text with as high a level of accuracy as possible. To accomplish this, we must:

1. Select a semantic tag-set to use
2. Apply that tag-set to Irish texts and compile a semantic lexicon for Irish using those tags
3. Decide how to measure the accuracy of the tagging components developed
4. Develop a pipeline to assign these tags, using whatever methods and resources necessary
5. Score the output of the pipeline using the accuracy measurement method

The remainder of section 3 and section 4 will describe these tasks in detail.

3.2 Semantic tag-set selection

There are a number of existing systems of semantic tags that can be used. It was decided to use the USAS annotation system for the following reasons. Firstly, the USAS system has an extensive set of tags that are encyclopaedic in their organisation. When selecting a tag for a particular word, it is always possible to find a category. Secondly, the USAS tags are arranged hierarchically, meaning that if a more specific category isn't applicable, there is always a broader category that applies. Additionally, this hierarchy allows for programmatic analysis and scoring. Thirdly, the USAS system is well integrated into the PyMUSAS library. It would be possible to use another tag-set with PyMUSAS, e.g. the framework proposed in Bick (2000), but it would require additional components to be built. This would be worth exploring in the future.

One drawback of using the USAS tag-set is that reading the tags themselves gives little indication of what category they represent. For example, the tag O1.2 represents the category *Substances and materials generally: Liquid*. This makes manual tagging and verification more costly, as tag definitions must be continually referenced.

3.3 Applying USAS to Irish text

For the purposes of this project, three texts were manually tagged, creating a 'golden standard' corpus of 'perfectly' tagged data

- A paragraph (110 tokens) from an online news article about a controversy concerning an Irish politician
- The first 226 tokens of the Irish Wikipedia article about the American TV show *The Wire*
- The first 301 tokens of the Irish Wikipedia article about the British author George Orwell

This corpus contains a total of 717 tokens. Each text was tagged by one or other of the paper's authors, and checked by the other. The texts were chosen as convenient examples of contemporary language. They are by no means representative of the variety in written and spoken registers of Irish, but they give enough data to drive standardisation of tagging conventions, and they provide many interesting cases of semantic ambiguity to drive innovation in semantic disambiguation. Even this small amount of manually tagged and checked text provides adequate data for testing the accuracy of the semantic tagging pipeline.

3.4 Semantic lexicon compilation

PyMUSAS is, at its core, a system for matching lemmas and part-of-speech tags, and assigning corresponding semantic tags. PyMUSAS has two types of lexicon format, the single-word lexicon and the multi-word lexicon. The single-word lexicon format is a tab separated value (TSV) file containing the fields *lemma, pos* and the *semantic_tags* field which contains a space-separated list of semantic tags to be assigned to any token in a text that matches that lemma and POS tag. The multi-word lexicon format is a TSV file containing the columns *mwe_template* and *semantic_tags*. The *mwe_template* field contains a pattern of words to match, and the *semantic_tags* field is identical to that of the single-word lexicon.

3.4.1 Lexicon derived from the golden standard

With the creation of the golden standard corpus, we have an initial dataset for creation of the single-word and multi-word PyMUSAS lexicons. This was done by taking each token or multi-word expression in the data and using the online version of the Ó Dónaill (1977) Irish-English dictionary to gather not only the sense which was used in the text itself, but all senses. Each sense was assigned a USAS tag, and these were added to the lexicon file in the order that the senses were listed in the dictionary. This lexicon is hereafter referred to as the *"Manually built"* lexicon. It should be noted that the sense-order found in the dictionary is not necessarily in descending order of likelihood. Dictionary-sense ordering in the past often started with the most prototypical sense of the word, without taking into account frequency of sense usage.

Building a lexicon from manually tagged data will give good coverage when tagging similar data, but in order to create a more comprehensive lexicon, other sources are necessary.

3.4.2 Lexicon based on the NEID

The New English-Irish Dictionary (NEID) hosted at www.focloir.ie (Ó Mianáin, 2013) contains roughly 80,000 English headwords. Each entry contains a list of senses of the English word, and their respective translations in common modern Irish usage. Many of the sense entries have one or more subject categories associated with them. For example, the senses of the headword *doctor* have the following subject categories associated with them: *empl* (Employment), *med* (Medical), and *ed* (Education). These semantic categories have the distinct advantage of having been manually checked and verified by lexicographers. Not every entry in the NEID has subject tags associated with its senses. Nevertheless, this is a substantial source of semantic category data.

In order to exploit this data for a USAS lexicon, the English headword, together with the Irish translation for each sense of each entry along with any subject tags associated were extracted. Those subject tags were then mapped to comparable USAS categories. Upon extraction of the data from NEID, 21,734 of the English headwords had both a single-word Irish translation and one or more associated subjects. Accounting for duplicate Irish translations (different English headwords can have the same Irish translation), 14,830 Irish lemmas were extracted, along with their part of speech and associated subjects. This lexicon is hereafter referred to as the *"NEID"* lexicon.

3.4.3 Lexicon based on English USAS lexicon

Another useful source of semantic information is the USAS single-word lexicon for English. This lexicon available with PyMUSAS contains 55,707 entries, each containing a list of possible USAS tags, though it is unconfirmed whether they are in descending order of likelihood. To make use of this lexicon, it is required to obtain direct single-word Irish translations of as many of the entries as possible. This was done using several sources:

- Firstly, a manually curated Irish translation of the "Core" WordNet list (Boyd-Graber et al., 2006) was used to add 3,923 Irish translations. "Core" WordNet is a list of 5,000 of the most frequently used word senses
- The NEID database was then used to automatically translate 1,139 additional entries.
- The Ó Dónaill (1977) English-Irish dictionary was next used to translate 2,769 additional entries.
- The de Bhaldraithe (1959) English-Irish dictionary was finally used to translate 2,333 additional entries.

This resulted in a total of 9,039 entries in the English USAS single-word lexicon having single-word Irish translations. Taking duplicates into account, this provided 7,960 Irish lemma/POS pairs with associated USAS tags. These entries were assembled into a USAS single-word lexicon for Irish. This lexicon will be hereafter referred to as the *"USAS-en"* lexicon.

3.4.4 Combined lexicon

Combining the NEID and USAS-en lexicons can potentially bring the best of both sources. The NEID lexicon has 14,830 entries, and the USAS-en lexicon has 7,960 entries. Of the lemma-pos pairs in these two lexicons 3,935 are common to both, meaning that their lists of possible USAS tags must be merged. Merging the entries for common lemma/POS pairs entailed checking whether each tag in the NEID list matched any tag in the USAS-en list, and if not adding it to the end of the list. The resultant combined single-word lexicon has a total of 18,866 entries. This lexicon will be hereafter referred to as the *"NEID + USAS-en"* lexicon.

3.5 The existing pipeline

Previous research by Uí Dhonnchadha and Van Genabith (2006) has produced a reusable pipeline for part-of-speech annotation of Irish. This pipeline uses finite-state transducers and constraint grammar to assign lemmas and part-of-speech tags in the PAROLE format (Uí Dhonnchadha, 2011). Dependency tagging adds syntax tree detection to the pipeline (Uí Dhonnchadha, 2009). This pipeline has been wrapped in a SpaCy component and added to a SpaCy pipeline, allowing semantic tagging components to be inserted after the existing components.

3.6 Applying PyMUSAS

The output from the existing pipeline is first converted to SpaCy format. This output contains the original token, a lemma, a PAROLE tag and dependency information, providing enough to commence semantic annotation.

The PyMUSAS lexicon format requires a POS tag and a lemma for matching. The existing PAROLE tags contain not just the part of speech, but also other grammatical attributes such as gender, number and case, and so the PAROLE tags are shortened so that they represent the broad part of speech, which is all that is required for matching.

PyMUSAS is then applied to assign each token a list of semantic tags from the lexicons. PyMUSAS first uses the single word lexicon to match all tokens. It then uses the multi-word lexicon to match multi-word patterns. In cases where a multi-word pattern is matched, the tags listed in the multi-word lexicon completely override any tags assigned from the single-word lexicon. Any tokens that aren't matched are assigned the tag `Z99` (Unmatched).

As described in section 2.1.1, when using the USAS tag-set, a token may be annotated with multiple space-separated tags, signifying a list of possible tag matches. The USAS guide (Archer et al., 2002) does not state the significance of the order of these tags, but it was decided that for this project we would treat it as the descending order of likelihood. This allows components of the tagging pipeline to make decisions based on this assumption. It is also useful when scoring the accuracy of the tagging pipeline, as discussed in section 4.

3.7 Semantic disambiguation

The greatest challenge of semantic annotation is disambiguating words with many senses. For example, the English word *"stock"* can have meanings that can be categorised as financial, industrial, culinary, and more. These may all have different USAS categories, and indeed the entry in the English USAS lexicon for the noun *"stock"* is `A9+ F1 S4 O2 I1.1`, i.e. five possible categories based on various senses. It's worth noting here that an entry for a particular word in the USAS lexicon may not have the same number of possible categories as its dictionary entry has senses, since some categories may cover multiple senses. When all possible USAS tags are assigned to a token, this leaves the task of disambiguating which of them is the "correct" one, given the surrounding context. The remainder of this section discusses the methods of disambiguation that have been applied and tested.

3.7.1 Disambiguation using document-level categories

After PyMUSAS is applied, a document will have some tokens with a single semantic tag (which we'll refer to as *single-match* tokens), and some tokens with multiple possible matches (which we'll refer to as *multi-match* tokens). Single-match tokens do not require any disambiguation, and we can assume that their semantic category is unambiguous. This allows us to use them to disambiguate the multi-match tokens. There are several levels at which we could attempt this. We could do it at the sentence or phrase level, using single-match tokens to disambiguate their multi-match neighbours. However, since a given document will usually have a topic or theme, it is likely that the USAS tags for the tokens in that document will align to a relatively small number of top-level USAS fields.

For example, one of the documents in the 'golden standard' corpus is a newspaper article about a controversy surrounding an Irish politician and how their election posters were funded. In this article, the noun *aire* is used, and this word can have several meanings:

- 'care' (e.g. *tabhair aire do* meaning 'take care of')
- 'heed' (e.g. *ar aire!* meaning 'attention!')
- 'minister' (i.e. a governmental minister)

The lexicon entry for *aire* is therefore `S8+ A1.3 G1.1`, covering cover these three senses with the categories `S8` (Helping/hindering), `A1.3` (General: Caution) and `G1.1` (Government etc.). Given the subject of the newspaper article, a reader can tell that the most likely sense is in fact *minister*, and so the correct category is therefore `G1.1`.

To achieve this disambiguation programmatically, first we count the number of single-match tokens in the document that fall into each top-level USAS field, and we find that the most frequent of the three fields is `G` (Government and the public domain), followed by `S` (Social actions, states and processes) and finally `A` (General and abstract terms). Next, we re-order the possible tag matches for the token *aire* based on these frequencies, and we get `G1.1 S8+ A1.3`, which places the correct tag first in the list.

This technique is applied generally by a custom SpaCy component, added after PyMUSAS. Note that this component ignores the top-level field `Z` (Names & Grammatical Words), because it is a 'catch-all' for closed-class words, proper nouns and unmatched tokens. Making decisions based on the frequency of `Z` tags would therefore have a negative effect.

3.7.2 Year detection

The lexicon used for PyMUSAS applies the `N1` (Numbers) category to any token containing a number. However, the correct tag for a date or a year is `T1.3` (Time: Period). Determining when a number is not just a number but part of a date is another form of disambiguation, and there are various criteria that can be used to detect dates. Proximity of numbers to the names of months would be a clear indicator, but many years are also listed without months, e.g. *"he was born in 1901"*. The simplest way to detect dates is to pick out numbers in a range that are commonly used as years, and 'predict' that they are probably dates.

A component was therefore added that finds any tokens that have the `N1` tag, and if the numerical value falls in the range 1000 to 2100, it predicts that it is likely a year by inserting a `T1.3` tag before the `N1` in the list of possible matches. Retaining the `N1` tag as a less likely match doesn't eliminate the possibility that the token could be just a plain number.

4. Evaluating the tagger

Although the semantic tagging components are part of a larger NLP pipeline, we must measure their accuracy in isolation. For this we must compare 'perfect' lemma+POS input data for the semantic tagging components, and compare it with expected output (semantic tags). This section describes the novel way in which that comparison was performed.

4.1 Finding good accuracy measures

Lexical coverage is a measure of the proportion of tokens that are 'covered' (i.e. assigned tags) by a tagging component, as used to measure other USAS taggers in Piao et al. (2004) and Löfberg et al. (2005). Lexical coverage is useful, but it is a fairly broad metric. For tokens that receive semantic tags, we also need some way to measure the **accuracy** of those tags, and be able to measure any improvements in accuracy made by disambiguation methods.

Semantic tags have an inherently softer 'correctness' property than other tag types. This is because a given token could be a member of one or more fields, categories or sub-categories, depending on the word sense being used and the overall subject and context of the document or speech act. The USAS tag-set is a hierarchical set of fields, categories and sub-categories, with the ability to assign multiple compound tags for multi-category membership, and with additional postfixed symbols to denote degrees, semantic gender and more. It's also possible for a tagger to assign a list of possible matching tags in descending order of likelihood. This means that a tag assignment is not simply correct or incorrect. Instead, accuracy is a continuum.

Therefore, a floating-point number between 0 and 1 is used for each token's accuracy value (dubbed the *match value*) and a system of formulae and multipliers has been developed for scoring each token, taking into account the nuances of semantic tagging. This system is described below.

4.2 Single-tag accuracy

Each USAS tag has a top-level field, one or more subdivisions of that field, and several optional post-fixed symbols with a variety of functions. When comparing two tags, they can fully match, or have some degree of partial match. Table 1 shows the category definitions for the 'Government & the Public Domain' field.

G	**GOVT. & THE PUBLIC DOMAIN**
`G1`	Government, Politics and elections
`G1.1`	Government etc.
`G1.2`	Politics
`G2`	Crime, law and order
`G2.1`	Crime, law and order: Law and order
`G2.2`	General ethics
`G3`	Warfare, defence and the army; weapons

Table 1: USAS category **G** (*Govt. & the Public Domain*)

If a token's ideal tag is `G1.1` (Government etc.) but it is tagged as `G1` (Government, politics & elections), this is not incorrect, but it is less correct than if it were tagged with `G1.1` exactly. It would be less correct again if that token were tagged as `G1.2`, and less correct again if tagged as `G3`, etc. If the field and subdivisions of two tags are equal, but the post-fixed symbols differ, they can be said to be very similar, but not equal. Taking category `A5.1` (Evaluation: good/bad) as an example, `A5.1+` and `A5.1-` would likely be used for a pair of antonyms, and though the words might be closely related, they are not semantically the same.

To account for this variation while keeping a simple scoring scheme, the match values shown in Table 2 were used when comparing single tags.

Match status	**Correct tag**	**Tagger output**	**Score**
Completely unequal	`A1.2.3+`	`X4.3fi`	0.0
Equal top-level field	`A1.2.3+`	`A3.2`	0.4
Equal field and first subdivision	`A1.2.3+`	`A1.3`	0.6
One tag a subdivision of the other	`A1.2.3+`	`A1f`	0.7
Equal field and all subdivisions	`A1.2.3+`	`A1.2.3`	0.8
Completely equal	`A1.2.3+`	`A1.2.3+`	1.0

Table 2: Base match values

4.3 Compound-tag accuracy

Each tag could be a *compound* tag, i.e. a composition of multiple tags separated by forward-slashes, indicating membership of multiple categories (e.g. `F1/S2mf`). It is necessary to have a system for comparing these compound tags with other compound (and non-compound) tags. This system must somehow incorporate both the proportion of constituent tags that match, and the individual match values for each tag matched. The scheme used when comparing two compound tags is as follows:

- compare every combination of component tags from each compound tag
- calculate the proportion of component tags from each compound tag that were non-zero matches
- calculate the mean value of all non-zero values found
- multiply the two values above together to produce the overall match value for the two compound tags

This can be expressed as a formula:

```
match_value = mean(all non-zero match values) *
        proportion of non-zero matches)
```

This means that the match value is higher with higher proportions of matching component tags, and vice versa.

4.4 Multi-tag accuracy

USAS allows a tagger to assign multiple, space-separated tags to a token, signifying a list of possible tag matches. As stated previously, we refer to these as *multi-tag assignments*, and we assume these possible tag matches to be in decreasing order of likelihood. The tagging pipeline begins by assigning to each token a space-separated list of tags for all possible matches, and it will then narrow that list down using disambiguation methods. If the tagger is unable to narrow it down to a single tag, the space-separated list may still contain fully- or partially-correct tags, and so the various scenarios should be scored as a continuum of accuracy.

The 'golden standard', manually-tagged test data has single tag assigned to each token. These tags are assigned based on the sense of that token within its surrounding context, and for the purposes of measuring accuracy, we can call these the *correct* tags. If a particular tag in a *multi-tag assignment* list fully or partially matches the *correct* tag, we multiply that tag's match value by a multiplier value. This multiplier value is based both on that tag's position within the list (`position`), and on how many tags are in the list (`num_tags`). The value is between 0 and 1, and is calculated as weighted sum of these two elements, as follows:

```
multiplier = (position_weight / position) +
        (num_tags_weight / num_tags)
```

The sum of `num_tags_weight` and `position_weight` must be 1. For the purposes of measuring tagger accuracy, it was decided that position is more important than the number of tags in the list, and so values of 0.7 and 0.3 were used for `position_weight` and `num_tags_weight` respectively. This gives us a multiplier value with a relatively even distribution, as illustrated in Table 3.

position	num_tags	multiplier
1	1	1.000
1	2	0.850
1	3	0.800
1	4	0.775
2	2	0.500
2	3	0.450
2	4	0.425
3	3	0.333
3	4	0.308
4	4	0.250

Table 3: Distribution of multiplier value

This formula makes the assumption that there is an even distribution of likelihood between tags within a given space-separated tag list. This wouldn't be the case for the most part, but it is a good enough estimation for our purposes.

There could also be multiple tags in a space-separated list that partially match the correct tag. In this case, the multipliers are individually applied to the match value of each partial match, and then the final match value is taken as the mean of all these results, as follows:

```
match_value = mean(multiplier1 * tag1_mv,
          multiplier2 * tag2_mv)
```

4.5 Document-level accuracy

Using the methods defined above, each tag assignment in a document will be scored using a float value between 0 and 1, denoting how accurate that assignment is, compared to its ideal 'correct' semantic tag. The accuracy of one or more documents can therefore be calculated as the mean of the accuracy values of all tokens within. This is the ***accuracy*** value that has been used in all tests run against the semantic tagging pipeline, as reported in section 5.

4.6 Accuracy of non-USAS tag-sets

The description in this section has outlined how the accuracy of semantic tag assignments was calculated when using the USAS tag-set, and using the guidelines outlined by the USAS system. However, this method of calculating semantic tag accuracy can be applied more generally. This scoring method can be used with other tag-sets and systems as follows:

- Whenever there is a hierarchical category system, comparing two single tags can use the method outlined in section 4.2.
- Whenever a token can be assigned a multi-category (*compound*) tag, comparing these tags can use the method outlined in section 4.3.
- Whenever there are multiple possible tags of descending likelihood, the match value can be calculated as outlined in section 4.4.

5. Results and analysis

5.1 Tests using the golden standard corpus

A number of tests were run on the semantic tagging portion of the pipeline in isolation, using the manually annotated 'golden standard' corpus. As described in section 3.3, this corpus contains 3 documents, containing 717 tokens. Because the manually annotated corpus was used, the part-of-speech tags and lemmas were known to be 100% correct, and so this provided correct input to the semantic annotation components, hereby testing its performance in isolation.

Four sets of tests were run in this way, each with a different single-word lexicon:

- The lexicon manually built from the 'golden standard' corpus *("Manually built")*
- The lexicon based on the English USAS lexicon (*"USAS-en"*)
- The lexicon based on the New English Irish Dictionary (*"NEID"*)
- The combined English USAS and NEID lexicon *("USAS-en + NEID")*

With each of these lexicons, accuracy of semantic annotation was measured for the following pipeline configurations:

- Just the PyMUSAS component
- The PyMUSAS component and the document-level disambiguation component
- The PyMUSAS component and the year-detection component
- The PyMUSAS component, the document-level disambiguation component and the year-detection component

For each test, the following measurements were made:

- **Lexical coverage**, i.e. the percentage of all tokens that had semantic tags assigned.
- **Correctness**, i.e. the percentage of tokens with a single, fully correct semantic tag. This was measured for all tokens, and for content-word tokens only.
- **Accuracy**, i.e. the overall semantic tag accuracy, calculated as described in section 4. This was measured for all tokens, and for content-word tokens only.

5.1.1 Lexical coverage across the lexicons

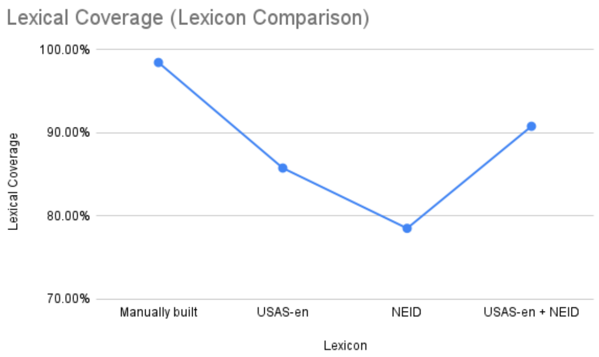

Figure 1: Lexical coverage across lexicons

Figure 1 shows lexical coverage across the four lexicons tested. For the manually-built lexicon, as would be expected, lexical coverage is 98.47%. The 2.53% gap represents several English words in the text that are not covered in the lexicon. As can be seen, the lexical coverage for the USAS-en and NEID lexicons are both considerably lower than 100%. But, at 85.77% and 78.52% respectively, their coverage is still substantial. However, when combined, they achieve a 90.8% lexical coverage. This indicates that there are many lexemes that are exclusive to one or the other of those sources.

It is notable that, even though it has almost double the number of lexical entries, the NEID lexicon has considerably less lexical coverage for these texts than the USAS-en lexicon. This might indicate that the USAS-en lexicon has better coverage of commonly used lexemes, due to its use of "Core" WordNet.

5.1.2 Correctness & accuracy across the lexicons

Figure 2 plots the correctness and accuracy values for all tokens and for content words across the four lexicons. As can be seen, correctness and accuracy using the externally-sourced lexicons are noticeably lower than when using the manually-built lexicon. This is expected, given that the manually built lexicon has close to 100% lexical coverage and has been created with the test data. Another observation is that the NEID lexicon, though it has almost double the number of lexical entries, scores lower in all measures than the USAS-en lexicon. As the lexical coverage of the NEID lexicon is lower than that of USAS-en, this is the most likely cause of the lower correctness and accuracy scores for this data source.

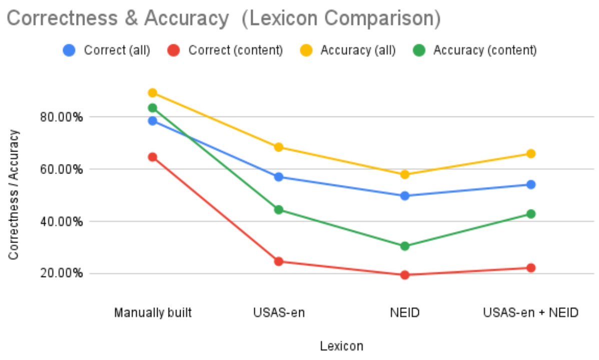

Figure 2: Correctness & accuracy across lexicons

The combined USAS-en+NEID lexicon scores marginally lower in both measures than the USAS-en does on its own, even though it achieves higher lexical coverage. This is most likely because the entries extracted from the NEID only contain broad categorical tags, and therefore don't increase the accuracy. Additionally, when combining the USAS-en and NEID lexicons, if an entry must be merged from both sources, the resultant semantic tag list can contain more potential matches, which can mean more ambiguity and therefore less accuracy.

5.1.3 Disambiguation effectiveness

Figure 3 is a plot of the overall accuracy for all tokens, across the various pipeline configurations (MUSAS only, MUSAS with the document-level disambiguator, MUSAS with the year detector, and all three components together). As can be seen, the manually built lexicon and the NEID lexicon achieve marginally increasing accuracy scores over the four pipeline configurations. However, this is not the case for the USAS-en or USAS-en + NEID lexicons. For these, there is a decrease in accuracy when document-level disambiguation is introduced.

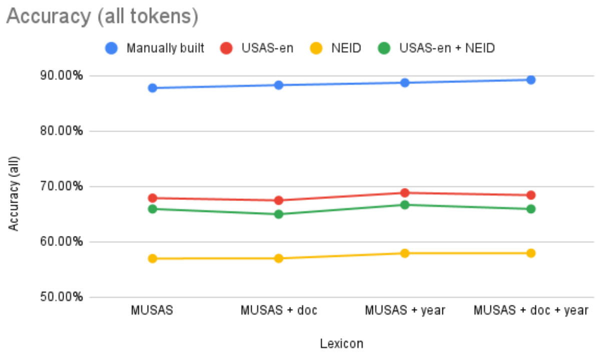

Figure 3: Accuracy across pipeline configurations

It is difficult to explain why this is, but, upon examination of the USAS-en lexicon entries, many of them contain large numbers of possible tags. Looking at an extreme example, *socraigh* is a verb that can mean 'settle', 'calm', or 'resolve'. The USAS-en lexicon entry for the word *socraigh* contains 17 possible tags from 10 different top-level fields, and many of them also compound tags. While the word *socraigh* may indeed be used in all of these ways, this is a challenging task for any disambiguation component. The document-level disambiguator is relatively simple. It counts the frequencies of top-level fields found in tokens that have been assigned a single tag, and then tries to re-order the tags in ambiguous tokens based on that. However, given the large number of lexicon entries in the USAS-en lexicon that contain large lists like this, the document-level disambiguator is likely not nuanced enough to handle them, and overall has a small negative effect on the accuracy score.

5.2 Testing on unseen data

Aside from testing against the golden standard corpus, a lexical coverage test was performed on a selection of 'unseen' articles, picked at (relative) random from Irish Wikipedia and two other news sites. This resulted in 1125 tokens of Irish text, which were processed through the existing NLP pipeline (Uí Dhonnchadha and Van Genabith, 2006), and then through the semantic annotation pipeline. The number of tokens matched by the semantic annotation components was measured to produce lexical coverage values, for each of the 4 lexicons, as plotted in figure 4.

The NEID lexicon scored the lowest for lexical coverage (70.76%), as in the golden standard tests. The Manually-built lexicon scored marginally higher (71.29%), but still quite low. This is to be expected since it is a small lexicon built from only the lexemes in the golden standard corpus. The USAS-en lexicon scored 77.51%, and the combined USAS-en + NEID lexicon scored 83.20%. There is more work to be done, but this is a promising level of coverage for a lexicon built from two incomplete sources.

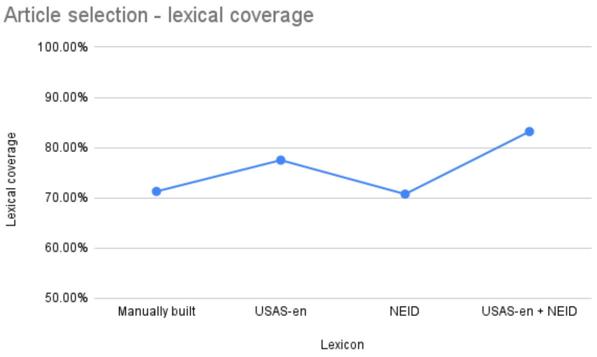

Figure 4: Lexical coverage for unseen data

6. Future research

While these initial results are promising, there is much more to be done.

The 'golden standard' corpus created for this project contained only 717 tokens from 3 documents. As discussed, these documents are not representative of the varied registers of spoken and written Irish, but the corpus was limited by the time available to manually annotate the texts. This corpus will be expanded by tagging and checking many more texts. As this is being done, the results can be used to expand the single and multi-word lexicons for the automatic tagger, thus increasing coverage.

The document-level disambiguation component and the year detector are good initial steps to resolving ambiguity, but there are other disambiguation methods that should be explored. For example, rather than just using document-level information, sentence-level or even clause level information could be used.

The MUSAS component currently tags any proper noun not contained within its lexicon as category Z0 (*Unmatched proper noun*). However, there exist more suitable categories for most proper nouns, including Z1 (*Personal names*), Z2 (*Geographical names*), Z3 (*Other proper names*) and M7 (*Places*). There are many rule-based methods that could be used to disambiguate these types of named entities, and further lists of personal, geographical and place names can be sourced.

Irish makes extensive use of phrasal verbs, prepositional verbs and modals involving the verb *bí 'to be'* and the copula *is* (used for states and emotions etc.). The resultant meaning of these constructions can often be different to the literal or prototypical meanings of the constituents. It would be possible to build a system to detect such constructions and alter semantic annotation accordingly.

Finally, once there is enough semantically annotated data via manual tagging or rule-based methods, and once it achieves an adequate level of accuracy, it could be used to train ML models. This could eventually be used to annotate Irish texts on-demand.

7. Conclusions

This project has taken some valuable steps towards a standard semantic tagging framework for Irish. It has explored conventions for tagging Irish using the USAS tag set, and produced a pipeline that achieves a good level of lexical coverage and a relatively high level of accuracy. It has also begun the process of constructing a standard semantic lexicon for Irish from various sources. There is still much research to be done to extend that pipeline, increase its accuracy, and extend the lexical resources, but it opens up opportunities to develop semantic role labelling, ML-based annotation, and much more.

This research into semantic annotation of Irish is part of the larger *Corpas Náisiúnta na Gaeilge* ('National Corpus of Irish') government-funded development project and all resources developed will be made available via the https://corpas.ie/ website. It is hoped that this research will contribute to the construction of widely available semantically annotated corpora of Irish, thereby unlocking benefits for further research and Irish language teaching.

8. Bibliographical references

Archer, D., Wilson, A. and Rayson, P. (2002), "Introduction to the USAS Category System." https://ucrel.lancs.ac.uk/usas/usas%20guide.pdf

Bick, E. (2000), "The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework."

Boyd-Graber, J., Fellbaum, C., Osherson, D. and Schapire, R. (2006), "Adding Dense, Weighted Connections to WordNet." In Proceedings of Third International Global WordNet Meeting, Jeju Island, Korea.

CSO (2011), "Torthaí Daonáirimh 2011." Technical report, Central Statistics Office.

Fellbaum, C. (1998), "WordNet: An Electronic Lexical Database". MIT Press.

Fillmore, C. J. (1968), "The Case for Case", 1–88. Holt, Rinehart & Winston.

Fillmore, C. J., Johnson, C. R. and Petruck, M. R. L. (2003), "Background to FrameNet." International Journal of Lexicography, 16, 235–250.

Gildea, D. and Jurafsky, D. (2002), "Automatic Labeling of Semantic Roles". In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong.

Givón, T. (1990), "Syntax: A Functional-Typological Introduction, volume 2". John Benjamins.

Hanks, P. (2004). "Corpus Pattern Analysis". Proceedings of Euralex. Lorient, France.

Hanks, P. (2013). "Lexical Analysis: Norms and Exploitations". Cambridge, MA: MIT Press.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020), "spaCy: Industrial-strength Natural Language Processing in Python." URL https://github.com/explosion/spaCy/blob/master/CITATION.cff.

Hovy, E. (2022), Corpus Annotation. In: The Oxford Handbook of Computational Linguistics. OUP.

Jackendoff, R. S. (1972), "Semantic Interpretation in

Generative Grammar". MIT Press.

Jackendoff, R. S. (1990), "Semantic Structures". MIT Press.

Kipper Schuler, K. (2005), "VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon." Dissertation Abstracts International, B: Sciences and Engineering, 66.

Koeva, S., Dimitrova, C., Stefanova, V. and Hristov, D. (2018). "Mapping WordNet Concepts with CPA Ontology". In Proceedings of the 9th Global Wordnet Conference. Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Li, T. and Srikumar, V. 2019. "Augmenting Neural Networks with First-order Logic". In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy.

Löfberg, L., Piao, S., Rayson, P., Juntunen, J-P., Nykänen, A. and Varantola, K. (2005), "A semantic tagger for the Finnish language."

Markchom, T., Liang, H., Gitau, J., Liu, Z., Ojha, V., Taylor, L., Bonnici, J. and Alshadadi, A. (2023). UoR-NCL at SemEval-2023 Task 1: Learning Word-Sense and Image Embeddings for Word Sense Disambiguation. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). Toronto, Canada. ACL.

Melamud, O., Goldberger, J. and Dagan, I. (2016). "context2vec: Learning Generic Context Embedding with Bidirectional LSTM". In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning Berlin, Germany

Miller, G. A. (1995), "WordNet: A Lexical Database for English." Communications of the ACM, 38, 39–41.

Palmer, F. R. (1981), "Semantics", 2nd edition. Cambridge University Press.

Palmer, M., Kingsbury, P., and Gildea, D. (2005), "The Proposition Bank: An Annotated Corpus of Semantic Roles." Computational Linguistics, 31, 71–106.

Piao, S., Bianchi, F., Dayrell, C. D'Egidio, A. and Rayson, P. (2015), "Development of the Multilingual Semantic Annotation System." In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez-Yáñez, R.M., Knight, D., Křen, M., Lofberg, L. Adeel Nawab, R.M., Shafi, J., Lee Teh, P., Mudraya, O. (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.

Piao, Scott, Paul Rayson, Dawn Archer, and Tony Mcenery (2004), "Evaluating Lexical Resources for A Semantic Tagger." In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal.

Piao, S., Rayson, P. and Watkins, G. (2017). "Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.

Pustejovsky, James and Amber Stubbs (2012), "Natural Language Annotation for Machine Learning". O'Reilly.

Rayson, P., Archer, D., Piao, S. and McEnery, T. (2004), "The UCREL Semantic Analysis System." URL http://mot.kielikone.fi/benedict/.

Scannell, K. P. and O'Regan, J. (2017), "History and Development of the Irish Language Semantic Network Special Issue on Linking, Integrating and Extending Wordnets." URL https://cadhan.com/lsg/.

Uí Dhonnchadha, E. (2009), "Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar." URL https://doras.dcu.ie/2349/.

Uí Dhonnchadha, E. (2011), "PAROLE Morphosyntactic Tagset for Irish." URL https://www.scss.tcd.ie/~uidhonne/parole.htm.

Uí Dhonnchadha, E. and Van Genabith, J. (2006), "A Part-of-speech Tagger for Irish Using Finite-State Morphology and Constraint Grammar Disambiguation." In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.

Wigger, A. (2008), "Advances in the lexicography of Modern Irish verbs." Issues in Celtic Linguistics. Lublin Studies in Celtic Languages 5, 233–250.

9. Language resource references

New Corpus for Ireland (2013), "New Corpus for Ireland." URL http://corpas.focloir.ie.

De Bhaldraithe, T. (1959) . English-Irish Dictionary. An Gúm. URL https://www.teanglann.ie/en/eid/

Ó Dónaill, N. (1977), Foclóir Gaeilge-Béarla. An Gúm. URL https://www.teanglann.ie/en/fgb/

Ó Mianáin, P. (2013), "New English-Irish Dictionary." URL https://www.focloir.ie/.