WNLT
Welsh Natural Language Toolkit

Introduction

The Welsh Natural Language Toolkit (WNLT) is a Welsh Government funded project under the Welsh-language technology and digital media grant. The toolkit contains a set of four core Natural Language Processing (NLP) modules that enable the development of generic computational linguistic applications and contribute to the Welsh language technology infrastructure a much needed open source NLP toolkit. The project builds on the General Architecture for Text Engineering (GATE) by adapting and expanding existing modules (plugins) to Welsh.

The Toolkit contains the following **four modules** (plugins)

- Tokenizer
- Sentence Splitter
- Part of Speech Tagger
- Morphological Analyser

The modules benefit from a combination of glossaries with algorithmic arrangements that address specific linguistic behaviours of the Welsh language.

Tokenizer

The WNLT Tokenizer extends the default GATE Tokenizer and similarly splits the text into very simple tokens such as numbers, symbols and words of different types. The Tokenizer distinguishes words in uppercase, lowercase, and between types of symbols. The module uses a slightly modified version of the original GATE Tokenizer rules file and an extended JAPE post-processing transducer adapting the generic output of the Tokenizer to the requirements of the Welsh part-of-speech tagger.

Token Types

The WNLT Tokenizer delivers the same types of Tokens and Space Tokens with default ANNIE Tokenizer as listed below:
- **[Word]** including the attribute 'orth' that takes the values; upperInitial, allCaps, lowerCase, mixedCaps
- **[Number]** any combination of consecutive digits.
- **[Symbol]** any special character is a symbol
- **[Space Token]** white spaces which are divided into two types of SpaceToken - space and control

Welsh Tokenizer Modifications

The Welsh Tokenizer uses a modified version of the GATE Tokenizer file *'AlternateTokeniser.rules'* which originally splits hyphenated and apostrophised cases into separate tokens. This behaviour is desirable due to the extensive and elaborate use of hyphens and apostrophe in Welsh which differs significantly from English, for example use of hyphens in adjectival compounds. A succeeding post-processing transducer joins under a single token several types of hyphenated and apostrophised constructs. The modified version also merges punctuation and symbol under a single Token type named *'symbol'*.

The modified post-processing transducer joins together in a single token **the following constructs**:

- Hyphenated placenames e.g. Llanarmon-yn-Ial
- Compounds of the common prefix e.g. ad-dala, cyd-ddefnyddir, rhag-glorineiddia
- Separate constituents hyphenation for the cases d+d, d+dd, dd+d, dd+dd, ff+f, ng+g, g+g, l+l, ll+l, t+h e.g. ladd-dy, cybydd-dod, cyd-dyfu, hwynt-hwy
- Apostrophe loss of vowel initialy e.g. 'Deryn
- Apostrophe loss of vowel medially eg. i'engoed
- Apostrophe loss of final consonant e.g. cry' for cryf hapusa' for hapusaf
- Apostrophe for common contractions, cases:i,m,n,r,w,ch,th
- Ordinals e.g. 1af, 2il, 3ydd, 4ydd
- Special cases of prepositions: Ar gyfer , Er mwyn , Yn erbyn, and Oddi followed by a preposition.

Init-time parameters

encoding
The character encoding to be used for reading the input

tokenizerRulesURL
The path to the Tokenizer rules files, the default file is located at */resources/Tokeniser/WelshTokeniser.rule*

transducerGrammarURL
The path to the post-processing tranducer grammar, the default JAPE file is located at */resources/Tokeniser/postprocess.jape*

Run-time parameters

annotationSetName
The name for annotation set where the resulting Token annotations will be created. It is optional, if left blank then the *'default'* annotation set is assigned.

Sentence Splitter

The WNLT sentence splitter segments the text into sentences using the same set of JAPE grammars used in ANNIE. Hence, it delivers annotations of type *'Sentence'* and *'Split'*. It also makes available an alternative ruleset (main-single-nl.jape), which considers newlines and carriage returns differently. The alternative ruleset, similarly to ANNIE, should be used when a new line on the page indicates a new sentence.

Sentence Splitter Modifications.

The WNLT sentence splitter uses a list of abbreviations adapted to Welsh that help distinguish sentence-marking full stops from other kinds. The abbreviations list contains 330 entries of the following categories:

1. Linguistic e.g. abs (absolute), cfst (synonym)
2. Narrative eg Brth (British) , e.e (for example)
3. Science e.g. Seic (Psychology), Tiwt (Teutonic)
4. Spatial e.g. Morg (Glamorgan)
5. Temporal e.g. C.C (B.C), Mer (Wednesday)

Init-time parameters

encoding
The character encoding to be used for reading the input

gazetteerListsURL
The path to the gazetteer list of abbreviations, the default list is located at

/resources/sentenceSplitter/gazetteer/lists.def

transducerURL
The path to tranducer grammar, the default JAPE file is located at */resources/sentenceSplitter/grammar/main-single-nl.jape*

Run-time parameters

inputASName
The name of the annotation set used for input. It is optional, if left blank then the *'default'* annotation set is assigned.

outputASName
The name of the output annotation set where the resulting Split and Sentence annotations will be created. It is optional, if left blank then the *'default'* annotation set is assigned.

Part of Speech Tagger

The WNLT POS tagger is a modified version of the ANNIE's Hepple tagger. The tagger produces a part-of-speech tag as an annotation on each word or symbol. The list of tags used by the tagger is found below. The tagger uses a default lexicon which is based on the Free (GPL) Dictionary Eurfa v3.0.

List of Taggs

CC - coordinating conjunction: e.g. a, ac, fel, fod
CD - cardinal number
DT - determiner: e.g. y, yr, 'r
IN - preposition: e.g. am, ap, mewn
INT - interrogative: e.g. beth, ble, sut etc.
JJ - adjective
JJR - adjective comparative
JJS - adjective superlative
NN - noun singular or mass
NNS - noun plural
NNP - proper noun singular
NNPS - proper noun plural
NNM - noun masculine
NNF - noun feminine
PDT - pre-determiner: preceding an article or possessive pronoun; e.g. ambell, prif, rhai etc.
PP - pronoun
RP - particle, such as; gor, mi, na, nac, ni, ni's
RB - adverb
UH - interjection, such as; eh, huh, nefi, sori etc
VB - verb, base form
VBD - verb past tens
VBDP - verb pluperfect
VBDI - verb imperfect
VBI - verb infinitive
VBF - verb future
PN - punctuation, such as ’'[](){} 、 –...-!.?`^`″""";\V/
SC - special characters, all other cases such as £$%* etc.

Part of Speech Tagger Modifications.

The WNLT POS tagger uses a lexicon of 168669 pairs of terms and tags originating from the Eurfa dictionary. A mapping exercise has mapped the original Eurfa tags (http://www.eurfa.org.uk/abbrevs.php) to ANNIE

Hepple tagger like tags. Major modifications applied on the original POSTagger and Lexicon classes for classifying Welsh input. The classes were extended to recognise linguistic evidence that support word classification of unknown words beyond the limits of the Eurfa dictionary.

Init-time parameters

encoding
The character encoding to be used for reading lexicons and rules

lexiconURL
The path to the lexicon of terms-tags pairs, the default lexicon is located at */resources/postag/lexicon*

rulesURL
The path to the ruleset file, the default ruleset file is located at */resources/postag/ruleset*

Run-time parameters

inputASName
The name of the annotation set used for input

outputASName
The name of the annotation set used for output. This is an optional parameter. If user does not provide any value, new annotations are created under the default annotation set.

baseTokenAnnotationType
The name of the annotation type that refers to Tokens in a document (run-time, default = Token)

baseSentenceAnnotationType
The name of the annotation type that refers to Sentences in a document (run-time, default = Sentence).

outputAnnotationType
POS tags are added as category features on the annotations of type 'outputAnnotationType' (run-time, default = Token)

posTagAllTokens
If set to false, only Tokens within each baseSentenceAnnotationType will be POS tagged (run-time, default = true).

FailOnMissingInputAnnotations
if set to false, the PR will not fail with an ExecutionException if no input Annotations are found and instead only log a single warning message per session and a debug message per document that has no input annotations (run-time, default = true).

Morphological Analyser (Lemmatizer)

The Morphological Analyser takes as input a tokenized GATE document. Considering one token and its part of speech tag, one at a time, it identifies its lemma, mutation form and in some cases an affix. These values are then added as features on the Token annotation. The WNLT Morphological Analyser has significantly extended the original GATE Morphological Analyser to address the linguistic behaviour of Welsh with regards to inflection and mutation. The tool uses regular expression rules, a Lexicon of term-lemma pairs, a Gazetteer and a post-processing JAPE transducer for validating mutation propositions. The tool allows users to add new rules or modify the existing resources on their requirements.

Morphological Analyser Modifications

The rule file *default.rul*, which is available under the */resources/morph* directory is modified for the Welsh alphabet. The file contains regular expressions that address regular and irregular forms of plural constructs.

More information on how to write these rules can be found in GATE user guide at https://gate.ac.uk/sale/tao/splitch23.html#sec:parsers:morpher:rules The tool uses a Lexicon of 168794 term-lemma pairs for providing known lemmas, a post-processing JAPE transducer for the identification of mutation forms focusing on contact mutations of Soft, Nasal and Aspirate type. The lemmatization process is as follows:

1. **Lexicon Lookup**, if is a known word provide lemma from lexicon and exit, else if unknown **proceed to 2**
2. **Regular Expressions rules**, resolve lemma using rules and in any case **proceed to 3**
3. **Post-processing Transducer**, identify cases of contact mutation based on contextual evidence. Propose new lemmas based on the contextual evidence and hard-coded Welsh language rules and **proceed to 4**
4. **Check the validity of the proposed lemmas** against a gazetteer of 168785 valid Welsh lemmas and 5885 Welsh place names. If lemmas validate set the new lemma and exit, else **proceed to 5**
5. **Revert invalid lemma** to original non-mutated lemma form.

Init-time parameters

caseSensitive
By default, all tokens under consideration are converted into lowercase to identify their lemma and affix. If the user selects caseSensitive to be true, words are no longer converted into lowercase

encoding
The character encoding to be used for reading lexicons and rules

gazetteerListsURL
The path to the gazetteer list of valid lemmas, the default list is located at *resources/morph/gazetteer/lists.def*

lexiconURL
The path to the lexicon of terms-lemma pairs, the default lexicon is located at */resources/morph/lexicon*

rulesFile
The path to the file containing the regular expression patterns, the default file is located at */resources/morph/default.rul*

transducerURL
The path to post-processing tranducer grammar responsible for identification and proposition of mutations, the default JAPE file is located at */resources/morph/grammar/postprocess.jape*

validationTransducerURL
The path to tranducer grammar responsible for validating proposed mutations against the gazetteer of valid lemmas, the default JAPE file is located at */resources/morph/grammar/validation-main.jape*

Run-time parameters

affixFeatureName
Name of the feature that should hold the affix value.

rootFeatureName
Name of the feature that should hold the root value.

annotationSetName
Name of the annotationSet that contains Tokens.

considerPOSTag
Each rule in the rule file might have a separate tag, which specifies which rule to consider with what part-of-speech tag. If this option is set to false, all rules are considered and matched with all words.

failOnMissingInputAnnotations
If set to true (the default) the PR will terminate with an Exception if none of the required input Annotations are found in a document. If set to false the PR will not terminate and instead log a single warning message per session and a debug message per document that has no input annotations.

CYMRIE

CYMRIE is an Information Extraction (Named Entity Recognition) system for Welsh. The name CYMRIE is a paraphrasis of GATE's Information Extraction system ANNIE (A Nearly-New Information Extraction System). CYMRIE adapts ANNIE to Welsh input using a modified version of the NE Transducer of ANNIE targeted at the requirements of the Welsh language, for example adjective – noun constructs. The system is using a wide range of Welsh gazetteer lists to support the task of Named Entity Recognition while it maintains some of the original lists with focus on person names and place names. CYMRIE does not currently include a co-reference resolution module

The default annotation types, features and possible values produced by CYMRIE the same used in ANNIE and are based on the original MUC entity types, and are as follows:

- Person
  - gender: male, female
- Location
  - locType: region, airport, city, country, county, province, other
- Organization
  - orgType: company, department, government, newspaper, team, other
- Money
- Percent
- Date
  - kind: date, time, dateTime
- Address
  - kind: email, url, phone, postcode, complete, ip, other
- Identifier
- Unknown

CYMRIE Gazetteer lists

welsh_assembly_members, Major Type:person_full, Minor Type:government
welsh_charities, Major Type:organization, Minor Type:charity
welsh_coastal, Major Type:location, Minor Type:coastal
welsh_counties, Major Type:location, Minor Type:county
welsh_countries, Major Type:location, Minor Type:country
welsh_country_adj, Major Type:country_adj, Minor Type:COUNTRYADJ
welsh_country_denonyms, Major Type:country_adj, Minor Type:
welsh_currency_unit, Major Type:currency_unit, Minor Type:post_amount
welsh_date_key, Major Type:date_key, Minor Type:
welsh_date_unit, Major Type:date_unit, Minor Type:
welsh_days, Major Type:date, Minor Type:day
welsh_departments, Major Type:organization, Minor Type:government
welsh_facility, Major Type:facility, Minor Type:building
welsh_facility_key, Major Type:facility_key, Minor Type:
welsh_facility_key_ext, Major Type:facility_key_ext, Minor Type:
welsh_festival, Major Type:date, Minor Type:festival
welsh_goverment, Major Type:organization, Minor Type:government
welsh_govern_key, Major Type:govern_key, Minor Type:
welsh_greeting, Major Type:greeting, Minor Type:
welsh_hour, Major Type:time, Minor Type:hour
welsh_ident_prekey, Major Type:ident_key, Minor Type:pre
welsh_jobtitles_cap, Major Type:jobtitle, Minor Type:

```
welsh_jobtitles_lower, Major Type:jobtitle, Minor Type:
welsh_jobtitles_sen, Major Type:jobtitle, Minor Type:
welsh_lakes, Major Type:location, Minor Type:lake
welsh_loc_generalkey, Major Type:loc_general_key, Minor Type:
welsh_loc_key, Major Type:loc_key, Minor Type:post
welsh_loc_prekey, Major Type:loc_key, Minor Type:pre
welsh_ministry, Major Type:organization, Minor Type:government
welsh_months, Major Type:date, Minor Type:month
welsh_mountains, Major Type:location, Minor Type:mountain
welsh_number_fold, Major Type:number_fold, Minor Type:
welsh_numbers, Major Type:number, Minor Type:
welsh_ordinals, Major Type:date, Minor Type:ordinal
welsh_org_base, Major Type:org_base, Minor Type:
welsh_org_key, Major Type:org_key, Minor Type:
welsh_org_pre, Major Type:org_pre, Minor Type:
welsh_parishes, Major Type:location, Minor Type:parish
welsh_percent, Major Type:percent, Minor Type:
welsh_person_female, Major Type:person_first, Minor Type:female
welsh_person_female_amb, Major Type:person_first, Minor Type:female
welsh_person_female_cap, Major Type:person_first, Minor Type:female
welsh_person_male, Major Type:person_first, Minor Type:male
welsh_person_male_cap, Major Type:person_first, Minor Type:male
welsh_phone_prefix, Major Type:phone_prefix, Minor Type:
welsh_placenames, Major Type:location, Minor Type:city
welsh_political_parties, Major Type:organization, Minor Type:government
welsh_radio_stations, Major Type:organization, Minor Type:
welsh_regions, Major Type:location, Minor Type:region
welsh_rivers, Major Type:location, Minor Type:river
welsh_sport, Major Type:sport, Minor Type:
welsh_stop, Major Type:stop, Minor Type:
welsh_terranean, Major Type:location, Minor Type:terrain
welsh_time, Major Type:time, Minor Type:absolute
welsh_time_ampm, Major Type:time, Minor Type:ampm
welsh_time_modifier, Major Type:time_modifier, Minor Type:
welsh_time_unit, Major Type:time_unit, Minor Type:
welsh_timeofday, Major Type:timeofday, Minor Type:
welsh_timezone, Major Type:timeofday, Minor Type:
welsh_title, Major Type:title, Minor Type:civilian
welsh_title_female, Major Type:title, Minor Type:female
welsh_title_male, Major Type:title, Minor Type:male
welsh_unitary_authorities, Major Type:location, Minor Type:unitary_authority
welsh_university_uk, Major Type:organization, Minor Type:university
welsh_water, Major Type:location, Minor Type:region
```