Cross-lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study

Teresa Lynn[1,2], **Jennifer Foster**[1], **Mark Dras**[2] and **Lamia Tounsi**[1]
[1]CNGL, School of Computing, Dublin City University, Ireland
[2]Department of Computing, Macquarie University, Sydney, Australia
[1]{tlynn,jfoster,ltounsi}@computing.dcu.ie
[2]{teresa.lynn,mark.dras}@mq.edu.au

Abstract

We present a study of cross-lingual direct transfer parsing for the Irish language. Firstly we discuss mapping of the annotation scheme of the Irish Dependency Treebank to a universal dependency scheme. We explain our dependency label mapping choices and the structural changes required in the Irish Dependency Treebank. We then experiment with the universally annotated treebanks of ten languages from four language family groups to assess which languages are the most useful for cross-lingual parsing of Irish by using these treebanks to train delexicalised parsing models which are then applied to sentences from the Irish Dependency Treebank. The best results are achieved when using Indonesian, a language from the Austronesian language family.

1 Introduction

Considerable efforts have been made over the past decade to develop natural language processing resources for the Irish language (Uí Dhonnchadha et al., 2003; Uí Dhonnchadha and van Genabith, 2006; Uí Dhonnchadha, 2009; Lynn et al., 2012a; Lynn et al., 2012b; Lynn et al., 2013). One such resource is the Irish Dependency Treebank (Lynn et al., 2012a) which contains just over 1000 gold standard dependency parse trees. These trees are labelled with deep syntactic information, marking grammatical roles such as subject, object, modifier, and coordinator. While a valuable resource, the treebank does not compare in size to similar resources of other languages.[1] The small size of the treebank affects the accuracy of any statistical parsing models learned from this treebank. Therefore, we would like to investigate whether training data from other languages can be successfully utilised to improve Irish parsing.

Cross-lingual transfer parsing involves training a parser on one language, and parsing data of another language. McDonald et al. (2011) describe two types of cross-lingual parsing, direct transfer parsing in which a delexicalised version of the source language treebank is used to train a parsing model which is then used to parse the target language, and a more complicated projected transfer approach in which the direct transfer approach is used to seed a parsing model which is then trained to obey source-target constraints learned from a parallel corpus. These experiments revealed that languages that were typologically similar were not necessarily the best source-target pairs, sometimes due to variations between their language-specific annotation schemes. In more recent work, however, McDonald et al. (2013) reported improved results on cross-lingual direct transfer parsing using a universal annotation scheme, to which six chosen treebanks are mapped for uniformity purposes. Underlying the experiments with this new annotation scheme is the universal part-of-speech (POS) tagset designed by Petrov et al. (2012). While their results confirm that parsers trained on data from languages in the same language group (e.g. Romance and Germanic) show the most accurate results, they also show that training data taken across language-groups also produces promising results. We attempt to apply the direct transfer approach with Irish as the target language.

The Irish language belongs to the Celtic branch of the Indo-European language family. The natural first step in cross-lingual parsing for Irish would be to look to those languages of the Celtic language

[1]For example, the Danish dependency treebank has 5,540 trees (Kromann, 2003); the Finnish dependency treebank has 15,126 trees (Haverinen et al., 2013)

Proceedings of the First Celtic Language Technology Workshop, pages 41–49,
Dublin, Ireland, August 23 2014.

group, i.e. Welsh, Scots Gaelic, Manx, Breton and Cornish, as a source of training data. However, these languages are just as, if not further, under-resourced. Thus, we attempt to use the languages of the universal dependency treebanks (McDonald et al., 2013).

The paper is organised as follows. In Section 2, we give an overview of the status of the Irish language and the Irish Dependency Treebank. Section 3 describes the mapping of the Irish Dependency Treebank's POS tagset (Uí Dhonnchadha and van Genabith, 2006) to that of Petrov et al. (2012), and the Irish Dependency Treebank annotation scheme (Lynn et al. (2012b)) to the Universal Dependency Scheme. Following that, in Section 4 we carry out cross-lingual direct transfer parsing experiments with ten harmonised treebanks to assess whether any of these languages are suitable for such parsing transfer for Irish. Section 5 summarises our work.

2 Irish Language and Treebank

Irish, a minority EU language, is the national and official language of Ireland. Despite this status, Irish is only spoken on a daily basis by a minority. As a Celtic language, Irish shares specific linguistic features with other Celtic languages, such as a VSO (verb-subject-object) word order and interesting morphological features such as inflected prepositions and initial mutations, for example.

Compared to other EU-official languages, Irish language technology is under-resourced, as highlighted by a recent study (Judge et al., 2012). In the area of morpho-syntactic processing, recent years have seen the development of a part-of-speech tagger (Uí Dhonnchadha and van Genabith, 2006), a morphological analyser (Uí Dhonnchadha et al., 2003), a shallow chunker (Uí Dhonnchadha, 2009), a dependency treebank (Lynn et al., 2012a; Lynn et al., 2012b) and statistical dependency parsing models for MaltParser (Nivre et al., 2006) and Mate parser (Bohnet, 2010) trained on this treebank (Lynn et al., 2013).

The annotation scheme for the Irish Dependency Treebank (Lynn et al., 2012b) was inspired by Lexical Functional Grammar (Bresnan, 2001) and has its roots in the dependency annotation scheme described by Çetinoğlu et al. (2010). It was extended and adapted to suit the linguistic characterisics of the Irish language. The final label set consists of 47 dependency labels, defining grammatical and functional relations between the words in a sentence. The label set is hierarchical in nature with labels such as `vparticle` (verb particle) and `vocparticle` (vocative particle), for example, representing more fine-grained versions of the `particle` label.

3 A universal dependency scheme for the Irish Dependency Treebank

In this section, we describe how a "universal" version of the Irish Dependency Treebank was created by mapping the original POS tags to universal POS tags and mapping the original dependency scheme to the universal dependency scheme. The result of this effort is an alternative version of the Irish Dependency Treebank which will be made available to the research community along with the original.

3.1 Mapping the Irish POS tagset to the Universal POS tagset

The Universal POS tagset (Petrov et al., 2012) has been designed to facilitate unsupervised and cross-lingual part-of-speech tagging and parsing research, by simplifying POS tagsets and unifying them across languages. The Irish Dependency Treebank was built upon a POS-tagged corpus developed by Uí Dhonnchadha and van Genabith (2006). The treebank's tagset contains both coarse- and fine-grained POS tags which we map to the Universal POS tags (e.g. Prop Noun → NOUN). Table 1 shows the mappings.

Most of the POS mappings made from the Irish POS tagset to the universal tagset are intuitive. However, some decisions require explanation.

Cop → VERB There are two verbs 'to be' in Irish: the substantive verb *bí* and the copula *is*. For that reason, the Irish POS tagset differentiates the copula by using the POS tag `Cop`. In Irish syntax literature, there is some discussion over its syntactic role, whether it is a verb or a linking particle. The role normally played is that of a linking element between a subject and a predicate. However, Lynn et al. (2012a)'s syntactic analysis of the copula is in line with that of Stenson (1981), regarding it as a verb. In addition, because the copula is often labelled in the Irish annotation scheme as the syntactic head of the matrix clause, we have chosen `VERB` as the most suitable mapping for this part of speech.

Part-of-speech (POS) mappings			
Universal	**Irish**	**Universal**	**Irish**
NOUN	Noun Noun, Pron Ref, Subst Subst, Verbal Noun, Prop Noun	ADP	Prep Deg, Prep Det, Prep Pron, Prep Simp, Prep Poss, Prep CmpdNoGen, Prep Cmpd, Prep Art, Pron Prep
PRON	Pron Pers, Pron Idf, Pron Q, Pron Dem	ADV	Adv Temp, Adv Loc, Adv Dir, Adv Q, Adv Its, Adv Gn
VERB	Cop Cop, Verb PastInd, Verb PresInd, Verb PresImp, Verb VI, Verb VT, Verb VTI, Verb PastImp, Verb Cond, Verb FutInd, Verb VD, Verb Imper	PRT	Part Vb, Part Sup, Part Inf, Part Pat, Part Voc, Part Ad, Part Deg, Part Comp, Part Rel, Part Num, Part Cp,
DET	Art Art, Det Det	NUM	Num Num
ADJ	Prop Adj, Verbal Adj, Adj Adj	X	Item Item, Abr Abr, CM CM, CU CU, CC CC, Unknown Unknown, Guess Abr, Itj Itj, Foreign Foreign,
CONJ	Conj Coord, Conj Subord	.	. . .. ? ? ! ! : : ? . Punct Punct

Table 1: Mapping of Irish Coarse and Fine-grained POS pairs (coarse fine) to Universal POS tagset.

Pron Prep → ADP *Pron Prep* is the Irish POS tag for pronominal prepositions, which are also referred to as prepositional pronouns. Characteristic of Celtic languages, they are prepositions inflected with their pronominal objects – compare, for example, ***le*** *mo chara* 'with my friend' with ***le****is* 'with him'. While the Irish POS labelling scheme labels them as pronouns in the first instance, our dependency labelling scheme treats the relationship between them and their syntactic heads as `obl` (obliques) or `padjunct` (prepositional adjuncts). Therefore, we map them to `ADP` (adpositions).

3.2 Mapping the Irish Dependency Scheme to the Universal Dependency Scheme

The departure point for the design of the Universal Dependency Annotation Scheme (McDonald et al., 2013) was the Stanford typed dependency scheme (de Marneffe and Manning, 2008), which was adapted based on a cross-lingual analysis of six languages: English, French, German, Korean, Spanish and Swedish. Existing English and Swedish treebanks were automatically mapped to the new universal scheme. The rest of the treebanks were developed manually to ensure consistency in annotation. The study also reports some structural changes (e.g. Swedish treebank coordination structures). [2]

There are 41 dependency relation labels to choose from in the universal annotation scheme[3]. McDonald et al. (2013) use all labels in the annotation of the German and English treebanks. The remaining languages use varying subsets of the label set. In our study we map the Irish dependency annotation scheme to 30 of the universal labels. The mappings are given in Table 2.

As with the POS mapping discussed in Section 3.1, mapping the Irish dependency scheme to the universal scheme was relatively straightforward, due in part, perhaps, to a similar level of granularity suggested by the similar label set sizes (Irish 47; standard universal 41). That said, there were significant considerations made in the mapping process, which involved some structural change in the treebank and the introduction of more specific analyses in the labelling scheme. These are discussed below.

3.2.1 Structural Differences

The following structural changes were made manually before the dependency labels were mapped to the universal scheme.

coordination The most significant structural change made to the treebank was an adjustment to the analysis of coordination. The original Irish Dependency Treebank subscribes to the LFG coordination analysis, where the coordinating conjunction (e.g. *agus* 'and') is the head, with the coordinates as its dependents, labelled `coord` (see Figure 1). The Universal Dependency Annotation scheme, on the

[2]There are two versions of the annotation scheme: the *standard* version (where copulas and adpositions are syntactic heads), and the *content-head* version which treats content words as syntactic heads. We are using the *standard* version for our study.

[3]The `vmod` label is used only in the content-head version.

Dependency Label Mappings			
Universal	**Irish**	**Universal**	**Irish**
root	top	*csubj*	csubj
acomp	adjpred, advpred, ppred	*dep*	for
adpcomp	N/A	*det*	det, det2, dem
adpmod	padjunct, obl, obl2, obl_ag	*dobj*	obj, vnobj, obj_q
adpobj	pobj	*mark*	subadjunct
advcl	N/A	*nmod*	addr, nadjunct
advmod	adjunct, advadjunct, quant, advadjunct_q	*nsubj*	subj, subj_q
amod	adjadjunct	*num*	N/A
appos	app	*p*	punctuation
attr	npred	*parataxis*	N/A
aux	toinfinitive	*poss*	poss
cc	N/A	*prt*	particle, vparticle, nparticle, advparticle, vocparticle, particlehead, cleftparticle, qparticle, aug
ccomp	comp	*rcmod*	relmod
compmod	nadjunct	*rel*	relparticle
conj	coord	*xcomp*	xcomp

Table 2: Mapping of Irish Dependency Annotation Scheme to Universal Dependency Annotation Scheme

other hand, uses right-adjunction, where the first coordinate is the head of the coordination, and the rest of the phrase is adjoined to the right, labelling coordinating conjunctions as `cc` and the following coordinates as `conj` (Figure 2).

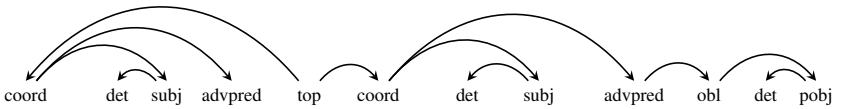

Figure 1: LFG-style coordination of original Irish Dependency Treebank

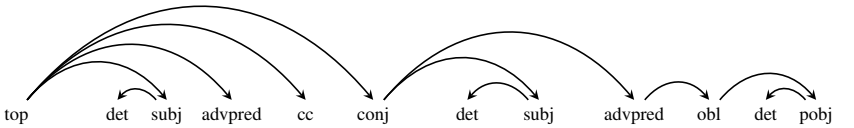

Figure 2: Stanford-style coordination changes to original Irish Dependency Treebank

subordinate clauses In the original Irish Dependency Treebank, the link between a matrix clause and its subordinate clause is similar to that of LFG: the subordinating conjunction (e.g. *mar* 'because', *nuair* 'when') is a `subadjunct` dependent of the matrix verb, and the head of the subordinate clause is a `comp` dependent of the subordinating conjunction (Figure 3). In contrast, the universal scheme is in line with the Stanford analysis of subordinate clauses, where the head of the clause is dependent on the matrix verb, and the subordinating conjunction is a dependent of the clause head (Figure 4).

3.2.2 Differences between dependency types

We found that the original Irish scheme makes distinctions that the universal scheme does not – this finer-grained information takes the form of the following Irish-specific dependency types: `advpred`,

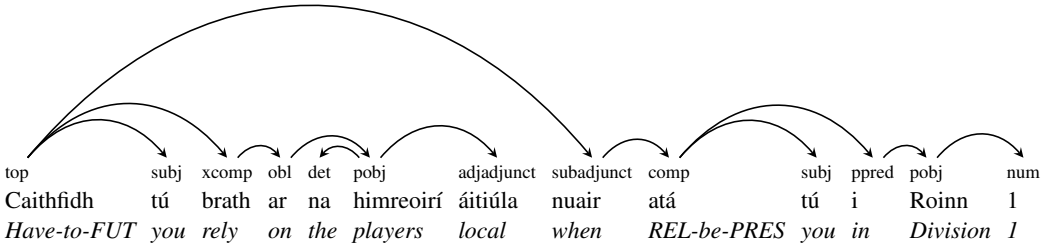

Figure 3: LFG-style subordinate clause analysis (with original Irish Dependency labels)

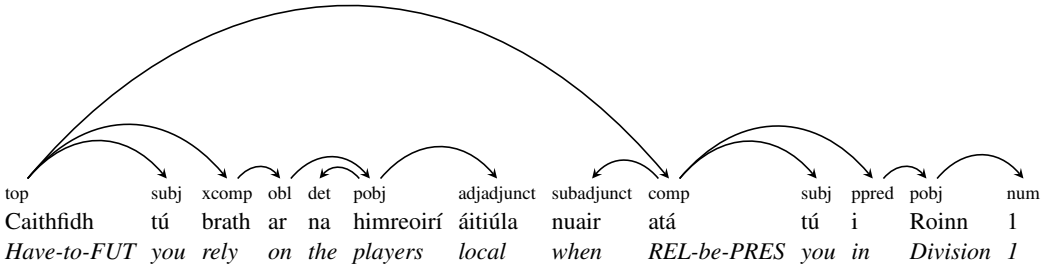

Figure 4: Stanford-style subordinate clause analysis (with original Irish Dependency labels)

ppred, subj_q, obj_q, advadjunct_q, obl, obl2. In producing the universal version of the treebank, these Irish-specific dependency types are mapped to less informative universal ones (see Table 2). Conversely, we found that the universal scheme makes distinctions that the Irish scheme does not. Some of these dependency types are not needed for Irish. For example, there is no indirect object iobj in Irish, nor is there a passive construction that would require nsubjpass, csubjpass or auxpass. Also, in the Irish Dependency Treebank, the copula is usually the root (top) or the head of a subordinate clause (e.g. comp) which renders the universal type cop redundant. Others that are not used are adp, expl, infmod, mwe, neg, partmod. However, we did identify some dependency relationships in the universal scheme that we introduce to the universal Irish Dependency Treebank (adpcomp, adposition, advcl, num, parataxis). These are explained below.

comp → adpcomp, advcl, parataxis, ccomp The following new mappings were previously subsumed by the Irish dependency label comp (complement clause). The mapping for *comp* has thus been split between *adpcomp*, *advcl*, *parataxis* and *ccomp*.

- adpcomp is a clausal complement of an adposition. An example from the English data is "some understanding of what the company's long-term horizon should **begin** to look like", where 'begin', as the head of the clause, is a dependent of the preposition 'of'. An example of how we use this label in Irish is: *an líne lántosach is mó clú a tháinig as Ciarraí ó **bhí** aimsir Sheehy ann* 'the most renowned forward line to come out of Kerry since Sheehy's time' (lit. 'from it was Sheehy's time'). The verb *bhí* 'was', head of the dependent clause, is an adcomp dependent of the preposition *ó*.

- advcl is used to identify adverbial clause modifiers. In the English data, they are often introduced by subordinating conjunctions such as 'when', 'because', 'although', 'after', 'however', etc. An example is "However, because the guaranteed circulation base is being **lowered**, ad rates will be higher". Here, 'lowered' is a advcl dependent of 'will'. An example of usage is: *Tá truailliú mór san áit mar nach **bhfuil** córas séarachais ann* 'There is a lot of pollution in the area because there **is** no sewerage system', where *bhfuil* 'is' is an advcl dependent of *Tá* 'is'.

- `parataxis` labels clausal structures that are separated from the previous clause with punctuation such as – ... : () ; and so on. Examples in Irish *Is léir go bhfuil ag éirí le feachtas an IDA – **meastar** gur in Éirinn a lonnaítear timpeall 30% de na hionaid* 'It is clear that the IDA campaign is succeeding – it **is believed** that 30% of the centres are based in Ireland'. Here, *meastar* 'is believed' is a `parataxis` dependent of *Is* 'is'.

- `ccomp` covers all other types of clausal complements. For example, in English, 'Mr. Amos says the Show-Crier team will probably **do** two live interviews a day'. The head of the complement clause here is 'do', which is a `comp` dependent of the matrix verb 'says'. A similar Irish example is: *Dúirt siad nach **bhfeiceann** siad an cineál seo chomh minic* 'They said that they don't **see** this type as often'. Here, *bhfeiceann* 'see' is the head of the complement clause, which is a `comp` dependent of the verb *Dúirt* 'Said'.

quant → num, advmod The Irish Dependency Scheme uses one dependency label (`quant`) to cover all types of numerals and quantifiers. We now use the universal scheme to differentiate between quantifiers such as *mórán* 'many' and numerals such as *fiche* 'twenty'.

nadjunct → nmod, compmod The Irish dependency label `nadjunct` accounts for all nominal modifiers. However, in order to map to the universal scheme, we discriminate two kinds: (i) nouns that modify nouns (usually genitive case in Irish) are mapped to `compmod` (e.g. *plean **margaíochta*** '**marketing** plan') and (ii) nouns that modify clauses are mapped to `nmod` (e.g. ***bliain** ó shin* 'a **year** ago').

4 Parsing Experiments

We now describe how we extend the direct transfer experiments described in McDonald et al. (2013) to Irish. In Section 4.1, we describe the datasets used in our experiments and explain the experimental design. In Section 4.2, we present the results, which we then discuss in Section 4.3.

4.1 Data and Experimental Setup

We present the datasets used in our experiments and explain how they are used. Irish is the target language for all our parsing experiments.

Universal Irish Dependency Treebank This is the universal version of the Irish Dependency Treebank which contains 1020 gold-standard trees, which have been mapped to the Universal POS tagset and Universal Dependency Annotation Scheme, as described in Section 3. In order to establish a monolingual baseline against which to compare our cross-lingual results, we perform a five-fold cross-validation by dividing the full data set into five non-overlapping training/test sets. We also test our cross-lingual models on an *delexicalised* version of this treebank.

Transfer source training data For our direct transfer cross-lingual parsing experiments, we use 10 of the standard version harmonised training data sets[4] made available by McDonald et al. (2013): Brazilian Portuguese (PT-BR), English (EN), French (FR), German (DE), Indonesian (ID), Italian (IT), Japanese (JA), Korean (KO), Spanish (ES) and Swedish (SV). For the purposes of uniformity, we select the first 4447 trees from each treebank – to match the number of trees in the smallest data set (Swedish). We delexicalise all treebanks and use the universal POS tags as both the coarse- and fine-grained values.[5] We train a parser on all 10 source data sets outlined and use each induced parsing model to parse and test on a *delexicalised* version of the Universal Irish Dependency Treebank.

Largest transfer source training data - Universal English Dependency Treebank English has the largest source training data set (sections 2-21 of the Wall Street Journal data in the Penn Treebank (Marcus et al., 1993) contains 39, 832 trees). As with the smaller transfer datasets, we delexicalise this dataset and use the universal POS tag values only. We experiment with this larger training set in order to establish whether more training data helps in a cross-lingual setting.

[4]Version 2 data sets downloaded from `https://code.google.com/p/uni-dep-tb/`

[5]Note that the downloaded treebanks had some fine-grained POS tags that were not used across all languages: e.g. VERB-VPRT (Spanish), CD (English).

Parser and Evaluation Metrics We use a transition-based dependency parsing system, MaltParser (Nivre et al., 2006) for all of our experiments. All our models are trained using the stacklazy algorithm, which can handle the non-projective trees present in the Irish data. In each case we report Labelled Attachment Score (LAS) and Unlabelled Attachment Score (UAS).[6]

4.2 Results

All cross-lingual results are presented in Table 3. Note that when we train and test on Irish (our monolingual baseline), we achieve an average accuracy of 78.54% (UAS) and 71.59% (LAS) over the five cross-validation runs. The cross-lingual results are substantially lower than this baseline. The LAS results range from 0.84 (JA) to 43.88 (ID) and the UAS from 16.74 (JA) to 61.69 (ID).

	SingleT										***MultiT***	***LargestT***
Training	EN	FR	DE	ID	IT	JA	KO	PT-BR	ES	SV	All	EN
UAS	51.72	56.84	49.21	**61.69**	50.98	16.74	18.02	57.31	57.00	49.95	57.69	51.59
LAS	35.03	37.91	33.04	**43.88**	37.98	0.84	9.35	42.13	41.94	34.02	41.38	33.97
Experiment	***SingleT-30***										***MultiT-30***	***LargestT-30***
Training	EN	FR	DE	ID	IT	JA	KO	PT-BR	ES	SV	All	EN
Avg sent len	23	24	16	21	21	9	11	24	26	14	19	23
UAS	55.97	60.98	53.42	**64.86**	54.47	16.88	19.27	60.47	60.53	54.40	61.40	55.54
LAS	38.42	41.44	36.24	**46.45**	40.56	1.19	10.08	45.04	45.23	37.76	44.63	37.08

Table 3: Multi-lingual transfer parsing results

A closer look at the single-source transfer parsing evaluation results (*SingleT*) shows that some language sources are particularly strong for parsing accuracy of certain labels. For example, `ROOT` (for Indonesian), `adpobj` (for French) and `amod` (for Spanish). In response to these varied results, we explore the possibility of combining the strengths of all the source languages (*multi-source direct transfer* (*MultiT*) – also implemented by McDonald et al. (2011)). A parser is trained on a concatenation of all the delexicalised source data described in Section 4.1 and tested on the full delexicalised Universal Irish Dependency Treebank. Combining all source data produces parsing results of 57.69% (UAS) and 41.38% (LAS), which is outperformed by the best individual source language model.

Parsing with the large English training set (*LargestT*) yielded results of 51.59 (UAS) and 33.97 (LAS) compared to a UAS/LAS of 51.72/35.05 for the smaller English training set. We investigated more closely why the larger training set did not improve performance by incrementally adding training sentences to the smaller set – none of these increments reveal any higher scores, suggesting that English is not a suitable source training language for Irish.

It is well known that sentence length has a negative effect on parsing accuracy. As noted in earlier experiments (Lynn et al., 2012b), the Irish Dependency Treebank contains some very long difficult-to-parse sentences (some legal text exceeds 300 tokens in length). The average sentence length is 27 tokens. By placing a 30-token limit on the Universal Irish Dependency Treebank we are left with 778 sentences, with an average sentence length of 14. We use this new 30-token-limit version of the Irish Dependency Treebank data to test our parsing models. The results are shown in the lower half of Table 3. Not surprisingly, the results rise substantially for all models.

4.3 Discussion

McDonald et al. (2013)'s single-source transfer parsing results show that languages within the same language groups make good source-target pairs. They also show reasonable accuracy of source-target pairing across language groups. For instance, the baseline when parsing French is 81.44 (UAS) and 73.37 (LAS), while the transfer results obtained using an English treebank are 70.14 (UAS) and 58.20(LAS). Our baseline parser for Irish yields results of 78.54 (UAS) and 71.59 (LAS), while Indonesian-Irish transfer results are 61.69 (UAS) and 43.88 (LAS).

The lowest scoring source language is Japanese. This parsing model's output shows less than 3% accuracy when identifying the `ROOT` label. This suggests the effect that the divergent word orders have

[6] All scores are micro-averaged.

on this type of cross-lingual parsing – VSO (Irish) vs SOV (Japanese). Another factor that is likely to be playing a role is the size of the Japanese sentences. The average sentence length in the Japanese training data is only 9 words, which means that this dataset is comparatively smaller than the others. It is also worth noting that the universal Japanese treebank uses only 15 of the 41 universal labels (the universal Irish treebank uses 30 of these labels).

As our best performing model (Indonesian) is an Austronesian language, we investigate why this language does better when compared to Indo-European languages. We compare the results obtained by the Indonesian parser with those of the English parser (*SingleT*). Firstly, we note that the Indonesian parser captures nominal modification much better than English, resulting in an increased precision-recall score of 60/67 on `compmod`. This highlights that the similarities in noun-noun modification between Irish and Indonesian helps cross-lingual parsing. In both languages the modifying noun directly follows the head noun, e.g. 'the statue of the hero' translates in Irish as *dealbh an laoich* (lit. statue the hero); in Indonesian as *patung palawan* (lit. statue hero). Secondly, our analysis shows that the English parser does not capture long-distance dependencies as well as the Indonesian parser. For example, we have observed an increased difference in precision-recall of 44%-44% on `mark`, 12%-17.88% on `cc` and 4%-23.17% on `rcmod` when training on Indonesian. Similar differences have also been observed when we compare with the French and English (*LargestT*) parsers. The Irish language allows for the use of multiple conjoined structures within a sentence and it appears that long-distance dependencies can affect cross-lingual parsing. Indeed, excluding very long sentences from the test set reveals substantial increases in precision-recall scores for labels such as `advcl`, `cc`, `conj` and `ccomp` – all of which are labels associated with long-distance dependencies.

With this study, we had hoped that we would be able to identify a way to bootstrap the development of the Irish Dependency Treebank and parser through the use of delexicalised treebanks annotated with the Universal Annotation Scheme. While the current treebank data might capture certain linguistic phenomena well, we expected that some cross-linguistic regularities could be taken advantage of. Although the best cross-lingual model failed to outperform the monolingual model, perhaps it might be possible to combine the strengths of the Indonesian and Irish treebanks? We performed 5-fold cross-validation on the combined Indonesian and Irish data sets. The results did not improve over the Irish model. We then analysed the extent of their complementarity by counting the number of sentences where the Indonesian model outperformed the Irish model. This happened in only 20 cases, suggesting that there is no benefit in using the Indonesian data over the Irish data nor in combining them at the sentence-level.

5 Conclusion and Future Work

In this paper, we have reported an implementation of cross-lingual direct transfer parsing of the Irish language. We have also presented and explained our mapping of the Irish Dependency Treebank to the Universal POS tagset and Universal Annotation Scheme. Our parsing results show that an Austronesian language surpasses Indo-European languages as source data for cross-lingual Irish parsing.

In extending this research, there are many interesting avenues which could be explored including the use of Irish as a source language for another Celtic language and experimenting with the projected transfer approach of McDonald et al. (2011).

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the CNGL (www.cngl.ie) at Dublin City University. We thank the three anonymous reviewers for their helpful feedback. We also thank Elaine Uí Dhonnchadha (Trinity College Dublin) and Brian Ó Raghallaigh (Fiontar, Dublin City University) for their linguistic advice.

References

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING'10*.

Joan Bresnan. 2001. *Lexical Functional Syntax*. Oxford: Blackwell.

Özlem Çetinoğlu, Jennifer Foster, Joakim Nivre, Deirdre Hogan, Aoife Cahill, and Josef van Genabith. 2010. LFG without C-structures. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Workshop on Crossframework and Cross-domain Parser Evaluation (COLING2008)*.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2013. Building the essential resources for Finnish: the Turku dependency treebank. *Language Resources and Evaluation*, pages 1–39.

John Judge, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha. 2012. *The Irish Language in the Digital Age*. Springer Publishing Company, Incorporated.

Matthias Kromann. 2003. The Danish Dependency Treebank and the DTAG Treebank Tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT2003)*.

Teresa Lynn, Özlem Çetinoğlu, Jennifer Foster, Elaine Uí Dhonnchadha, Mark Dras, and Josef van Genabith. 2012a. Irish treebanking and parsing: A preliminary evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1939–1946.

Teresa Lynn, Jennifer Foster, Mark Dras, and Elaine Uí Dhonnchadha. 2012b. Active learning and the Irish treebank. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 23–32.

Teresa Lynn, Jennifer Foster, and Mark Dras. 2013. Working with a small dataset – semi-supervised dependency parsing for Irish. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–11, Seattle, Washington, USA, October. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu, and Castelló Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL '13*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Nancy Stenson. 1981. *Studies in Irish Syntax*. Tübingen: Gunter Narr Verlag.

Elaine Uí Dhonnchadha and Josef van Genabith. 2006. A part-of-speech tagger for Irish using finite-state morphology and constraint grammar disambiguation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Elaine Uí Dhonnchadha, Caoilfhionn Nic Pháidín, and Josef van Genabith. 2003. Design, implementation and evaluation of an inflectional morphology finite state transducer for Irish. *Machine Translation*, 18:173–193.

Elaine Uí Dhonnchadha. 2009. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. Ph.D. thesis, Dublin City University.

CLTW 2014

The First Celtic Language Technology Workshop

Proceedings of the Workshop

A Workshop of the 25th International Conference on Computational Linguistics (COLING 2014) August 23, 2014 Dublin, Ireland

ISBN 978-1-873769-32-4
Proceedings of the Celtic Language Technology Workshop (CLTW)
John Judge, Teresa Lynn, Monica Ward and Brian Ó Raghallaigh (eds.)

Introduction

Language Technology and Computational Linguistics research innovations in recent years have given us a great deal of modern language processing tools and resources for many languages. Basic language tools like spell and grammar checkers through to interactive systems like Siri, as well as resources like the Trillion Word Corpus, all fit together to produce products and services which enhance our daily lives.

Until relatively recently, languages with smaller numbers of speakers have largely not benefited from attention in this field. However, modern techniques in the field are making it easier to create language tools and resources from fewer resources in a faster time. In this light, many lesser spoken languages are making their way into the digital age through the provision of language technologies and resources.

The Celtic Language Technology Workshop (CLTW) series of workshops provides a forum for researchers interested in developing NLP (Natural Language Processing) resources and technologies for Celtic languages. As Celtic languages are under-resourced, our goal is to encourage collaboration and communication between researchers working on language technologies and resources for Celtic languages.

Welcome to the First Celtic Language Technology Workshop. We received 15 submissions, and after a rigorous review process, accepted 12 papers. Eight of which will be presented as oral presentations and 4 of which will be presented at the poster session.