IRIS: English-Irish Machine Translation System

Mihael Arcan[†], Caoilfhionn Lane[†], Eoin Ó Droighneáin[‡], Paul Buitelaar[†]

[†]Insight Centre for Data Analytics, National University of Ireland, Galway
[firstname.lastname]@insight-centre.org

[‡]Acadamh na hOllscolaíochta Gaeilge, National University of Ireland, Galway
eoin.odroighneain@oegaillimh.ie

Abstract

We describe IRIS, a statistical machine translation (SMT) system for translating from English into Irish and vice versa. Since Irish is considered an under-resourced language with a limited amount of machine-readable text, building a machine translation system that produces reasonable translations is rather challenging. As translation is a difficult task, current research in SMT focuses on obtaining statistics either from a large amount of parallel, monolingual or other multilingual resources. Nevertheless, we collected available English-Irish data and developed an SMT system aimed at supporting human translators and enabling cross-lingual language technology tasks.

Keywords: Statistical Machine Translation, Under-resourced languages, Irish language

1. Introduction

The META-NET Language White Paper Series[1] showed that English has the best language technology support amongst all European languages, followed by languages such as Dutch, French, German, Italian and Spanish with moderate support. The same study showed that translations from moderately supported languages, such as French or Spanish, into English can achieve acceptable quality for many practical applications. Due to the lack of good translation resources, the Irish language has been categorised as a weak or not supported language by the META-NET report (Judge et al., 2012).

Despite the limited amount of resources, the demonstrated English-Irish SMT system IRIS[2] can nevertheless support human translators and enable information access across Irish and English language and culture. Additionally, we believe that through an SMT system, the broader community can gain access to information that might have otherwise been unavailable.

2. Related Work

Past research on translation systems for under-resourced languages focused on improving translation quality either by collaboratively collecting or generating parallel data needed for SMT models. Additionally, research focused on using closely related languages for translation improvement or using a pivot language to overcome the data sparseness.

With the aim of language preservation, Lewis and Yang (2012) show an SMT system for English to and from *White Hmong* (Hmong-Mien language). They built their system from dictionary entries and translations of introductions or phrases from localisation projects. To extend their parallel resources, they manually searched for Hmong phrases and its translations on the web, whereby they collected around 45,000 parallel sentences in overall. A different deployment of SMT systems in an under-resourced scenario was shown in Lewis et al. (2011) and Lewis (2010) as a consequence of the earthquake crisis in Haiti supporting emergency responders to find trapped people.

Differently, Babych et al. (2007) compare results between a direct transfer of an SMT system (source→target language) and translations via a cognate language (source→pivot→target language). Their approach focused on Slavic languages with Russian as the pivot language. The results showed the efficiency of the usage of dictionaries, grammars as well as lexical and syntactic similarities of closely related languages for translation improvements.

An early work dealing with translating Irish language was shown in Scannell (2006). The rule-based system was developed for translations of closely related languages, Irish (Gaeilge) and Scottish Gaelic (Gàidhlig), respectively. The translation system is based on a bilingual lexicon, which performs part-of-speech tagging, word sense disambiguation and a syntactic/lexical transfer. This work was expanded in Scannell (2014), focusing on overcoming the orthographical differences between the languages. As an additional task, the author casts the text normalisation problem as an SMT problem and applies the statistical models for normalisation of historical Irish text. The most recent work on a domain-specific English-Irish SMT is shown in Dowling et al. (2015), aiming to help Irish government with their translation tasks.

3. Irish language

Irish is a VSO language on the Celtic branch of the Indo-European language family tree. It is a highly inflected lan-

[1] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

[2] http://server1.nlp.insight-centre.org/iris/

guage, and the beginning, middle or end of a word may alter depending on the grammatical rule in question. There are four noun cases in the Irish language: the combined nominative and accusative case (*Cheannaigh sé an bord*, 'he bought the table'); the genitive case, which is used to indicate a number of noun relationships (*mata bo**i**rd*, 'table mat'; *ag glanadh an b**h**o**i**rd*, 'cleaning the table'); the dative case, which occurs when a noun is preceded by certain prepositions (*ar an **m**bord*, 'on the table'); the vocative case, which involves a form of address (*Seán*, a male personal name; *A S**h**eá**i**n*).

Other features of the language include feminine (*an aiste*, 'the essay'; *an b**h**ean*, 'the woman') and masculine (*an **t-**arásán*, 'the apartment'; *an bád*, 'the boat') grammatical genders, and numerous plural forms (*-a, -(a)igh, -(e)anna, -e, -ta, -(a)í*, etc.). A copula construction is used to indicate a permanent state (*Is dochtúir í*, 'She is a doctor'), while the substantive verb *bí* ('to be') is used for transient states (*Tá an ghriain ag taitneamh*, 'The sun is shining'). The language has habitual present and past tenses (*bím ag léamh*, 'I am (habitually) reading'; *bhínn i gcónaí ag léamh*, 'I was always reading'), and there are differing systems for cardinal, personal and ordinal numbers (*a dó*, '2'; *dhá b**h**ord*, 'two tables'; *beirt*, 'two people'; *an dara háit, 'second place'*) (Mac Congáil, 2004).

While Irish is the first official language of the Republic of Ireland, it is, perhaps paradoxically, a minority language threatened by language shift (Fishman, 1991). Native speakers of the language are mostly located in small, geographically separated, rural areas collectively known as the Gaeltacht. Traditionally, the language is classified into three spoken dialects (Connaught, Munster and Ulster), corresponding to the Gaeltacht areas in which they are spoken. Within the main dialects however, the language is even more diverse, with sub-dialects spoken by individual language communities. Moreover, the question of whether urban Irish should be considered a genuine dialect has recently received attention (Ó Broin, 2014).

The existence of distinct spoken dialects has had an effect on corpus planning in Ireland. As a compromise between the dialects, an artificial standard form of the language was developed for writing (Tulloch, 2006). This official standard was published in the middle of the last century, and included grammar recommendations along with spelling and orthographical reform (Rannóg an Aistriucháin, 1945; Rannóg an Aistriucháin, 1958). In a recent attempt to bring the standard closer to the spoken language, an updated standard was published, in which a wider range of grammatical variations are allowed (Uíbh Eachach, 2012).

The aforementioned corpus planning has complicated the development of language resources and tools. A large amount of quality texts written by native speakers before standardization are unusable for tasks such as language modelling unless they are converted to the standard form. Furthermore, consideration must be given to how much of the dialectal features are preserved in such a standardization process (Scannell, 2014). For example, it may be desirable to standardize spelling and orthography, but to preserve dialectal vocabulary and grammar. This is just one of many factors limiting the development of Irish language resources and tools for computational linguistics, and therefore the language has been defined as a less-resourced language (Piotrowski, 2012) in this domain.

4. English-Irish Machine Translation Development

Here we present IRIS, an English-Irish translation system, which is based on a widely used phrase-based SMT framework (Koehn et al., 2003). For generating the translation models, we use the statistical translation toolkit Moses (Koehn et al., 2007). Word alignments were built with GIZA++ (Och and Ney, 2003) and a 5-gram language model was built with KenLM (Heafield, 2011).

The monolingual and parallel corpora described in Section 4.2. are progressively added to the IRIS training set. This allows us to evaluate the performance of the system (Section 5.) at each point new data is added.

4.1. IRIS Framework

IRIS' bilingual interface (Figure 1) allows the user to enter English or Irish sentences that are to be translated into the target language. It also provides information on the current translation performance of IRIS in terms of the evaluation metric BLEU. Furthermore, it gives detailed information about the used data for the translation models accessed by IRIS. Finally, the system allows users to upload new parallel or monolingual (for language modelling) data into the system. If the uploaded data is recognised as monolingual or aligned parallel data in the targeted languages, the training step updates the translation and/or language models of IRIS accordingly to the new data. Lastly, a Web service provides access to translations for other applications beyond the graphical interface.[3]

4.2. English-Irish monolingual and parallel data

As shown in Table 1, IRIS currently[4] accesses translation models trained on around 1 million English-Irish parallel sentences. In addition to the monolingual data in parallel corpora, the language models are further enriched with additional monolingual data. We used 3.7 Million sentence of the News-2007 monolingual corpus[5] for the English language model and around 250,000 sentences extracted from the Wikipedia articles in Irish to enrich the Irish language model.

The majority of the parallel data used for the IRIS system is provided by the European Union institutions, which released a large number of multilingual resources in recent

[3] http://server1.nlp.insight-centre.org/iris/rest_service.html

[4] March 2016

[5] http://www.statmt.org/wmt15/translation-task.html

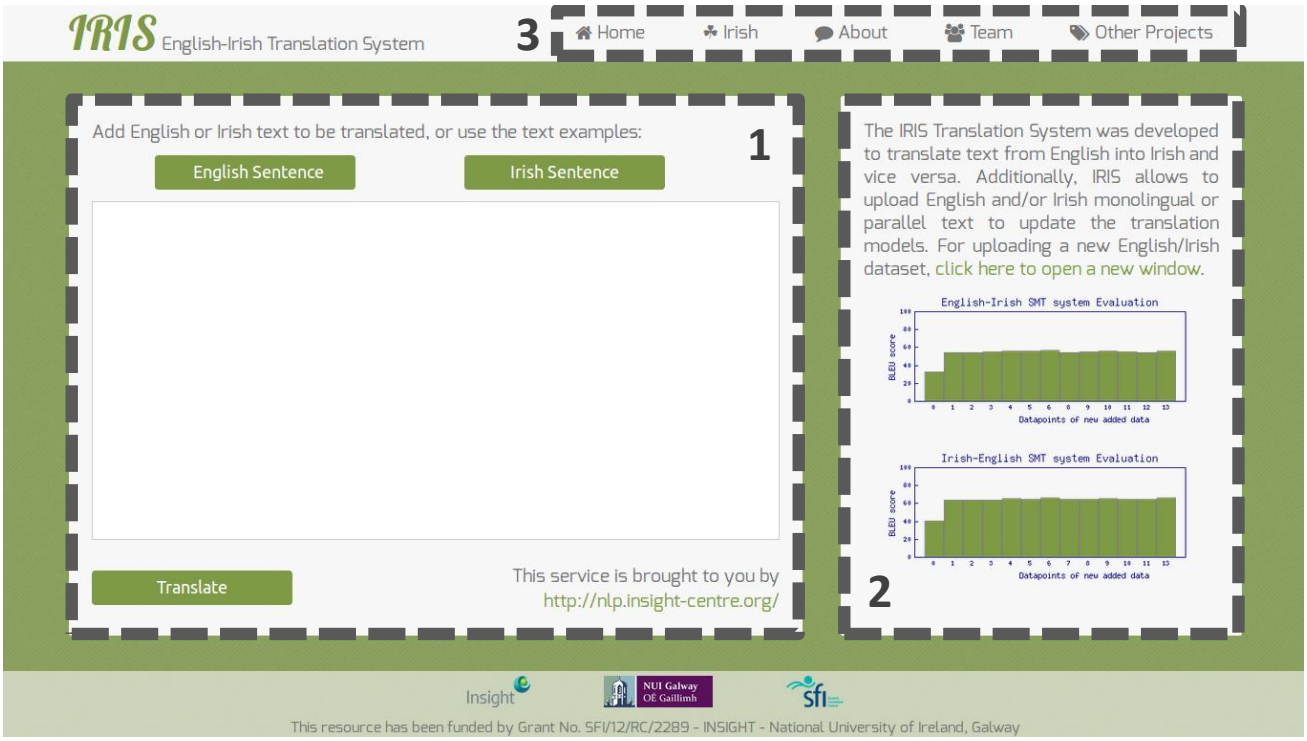

Figure 1: The graphical interface of IRIS, with the input option (1), system performance and data statistics (2) and additional information about the system (3).

Corpus	# lines	# English words	# Irish words
DGT	36,275	864,373	950,500
EU Bookshop	121,042	2,606,607	2,704,091
EU constitution	6,267	125,553	126,355
Focal$^{\diamond}$	213,683	414,730	440,228
GNOME$^{\diamond}$	75,051	288,916	297,882
Irish legislation	132,314	2,691,928	2,792,595
KDE4$^{\diamond}$	110,138	439,273	523,614
News-2007 (English)	3,782,548	90,490,396	/
Ubuntu$^{\diamond}$	191	1,038	1,103
Wikipedia Titles$^{\diamond}$	17,421	35,165	36,760
Irish sent. bank	3,895	31,655	32,800
Food and Beverages$^{\diamond}$	339	696	712
Wikipedia (Irish)	246,290	/	4,047,229
Textbooks	373,401	5,929,635	6,568,295
Apertium$^{\diamond}$	720	804	791
total (parallel)	1,096,117	13,573,234	14,684,356

Table 1: Statistics of the English-Irish data sets for SMT training (dictionary/terminological resources are marked with $\diamond$).

years. The *DGT* data set consists of translation memories generated by the *Directorate-General for Translation* in the European Commission. Skadiņš et al. (2014) collected multilingual documents from the *EU Bookshop* online platform,[6] which archives publications from various European institutions. The *European constitution* and its translation into Irish were used as well. We further included bilingual data from various localisation projects, i.e. *KDE4*, *GNOME*, *Ubuntu*. Differently to the aforementioned resources provided by the EU, these resources do not store aligned sentences, but rather store word to word translations, similar to dictionaries. All these resources were collected from the OPUS webpage[7] (Tiedemann, 2012).

We integrated further the translated *Irish legislations* parallel text and the English-Irish terminological database *Focal*, both provided by the Gaois platform.[8] In addition to the monolingual data extracted from Irish Wikipedia articles for the language model enhancement, we use the interlinked titles of Wikipedia to enrich the translation models. Although the dictionary or terminological knowledge used in IRIS represents less than 10% of bilingual knowledge (on word level) of our parallel data set, it holds valuable specific bilingual vocabulary, which is often not represented in parallel corpora. Also we used a data set of English-Irish sentences, i.e. *Irish Sentence Bank*[9] and extracted bilingual knowledge of food and beverages of the targeted languages.[10] A small English-Irish dataset was collected from the Apertium project,[11] a free/open-source platform for developing rule-based machine translation systems.

[6] http://bookshop.europa.eu/

[7] http://opus.lingfil.uu.se/

[8] http://www.gaois.ie/en/

[9] http://www.lexiconista.com/datasets/sentencebank-ga/

[10] http://www.gaeilge.ie/wp-content/uploads/2014/12/lamhleabhar_bia.pdf

[11] http://wiki.apertium.org/wiki/Main_Page

#	Corpus	English→Irish			Irish→English		
		BLEU	METEOR	chrF	BLEU	METEOR	chrF
0	DGT*	32.39	28.45	56.75	40.50	34.22	56.43
1	+EU Bookshop*	54.54	39.03	67.82	64.05	44.15	69.41
2	+EU constitution*	53.97	38.63	67.21	63.73	44.27	69.89
3	+Focal	54.82	39.30	68.59	64.04	44.65	70.53
4	+GNOME*	55.62	40.11	68.76	65.43	45.24	70.50
5	+Irish legislation	55.77	40.01	68.62	65.02	45.38	70.81
6	+KDE4*	56.62	40.73	69.36	66.28	46.24	71.23
7	+News-2007 (mono. English)	/	/	/	64.45	44.89	69.88
8	+Ubuntu	54.67	39.60	68.06	64.72	45.23	70.24
9	+Wikipedia Titles	55.44	39.99	68.10	65.41	45.52	70.55
10	+Irish sent. bank	55.76	40.23	68.35	64.87	45.72	70.92
11	+Food and Beverages	54.83	39.57	67.41	65.02	45.21	70.05
12	+Wikipedia (mono. Irish)	54.47	39.34	66.88	/	/	/
13	+Textbooks	55.84	40.28	68.61	66.18	46.12	71.21
14	+Apertium	54.85	39.66	67.75	64.68	45.06	70.11
	Google Translate	40.07	33.23	65.93	46.77	39.20	68.83

Table 2: Automatic translation evaluation based on BLEU, METEOR and chrF on different parallel/monolingual corpora (the evaluation data set was extracted from corpora annotated with *).

In addition to the publicly available parallel corpora, the Acadamh na hOllscolaíochta Gaeilge[12] at the National University of Ireland, Galway (NUIG) provided us with translations of second level textbooks (Cuimhne na dTéacsleabhar) in the domain of economics and geography. The data resource, funded by An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (COGG), holds around 350,000 parallel sentences or 6M English and 6.5M Irish words, respectively.

5. Evaluation

Here, we report results based on the BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and the chrF3 (Popović, 2015) metric for automatic evaluation of translations. Additionally, we perform a manual evaluation of the translations into Irish.

BLEU is calculated for individual translated segments (n-grams) by comparing them with a dataset of reference translations. Those scores, between 0 and 100 (perfect translation), are then averaged over the whole evaluation dataset to reach an estimate of the translation's overall quality. METEOR is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with standard exact word (or phrase) matching it has additional features, i.e. stemming, paraphrasing and synonymy matching. chrF3 is tokenisation-independent metric which has shown very good correlations with human judgements on the WMT2015 shared metric task (Stanojević et al., 2015), both on the system level as well as on the segment level, especially for morphologically rich(er) languages.

[12] http://www.acadamh.ie/index_irish.html

5.1. Automatic Evaluation

Our automatic evaluation is based on 2,000 parallel sentences randomly extracted from the *DGT*, *EUbookshop*, *EU constitution*, *GNOME* and *KDE4* corpora.[13]

Table 2 shows the evaluation of IRIS based for the progressively added monolingual and parallel data.[14] As seen, the BLEU score, which is calculated based on the overlap of the automatically generated translations and reference translations, improves as more data (monolingual or bilingual) is added to the system. Similarly to other experiments, translating into English performs slightly better than to translations into Irish, since English is less inflectional than the Irish language. Furthermore, based on our evaluation data set we outperform Google Translate[15] when translating to and from Irish.

5.2. Manual Evaluation of Irish Diploma data set

Additionally to the data set randomly selected from the aforementioned corpora, we used a small data set from the Translation and Interpreting Unit at NUIG (Irish Diploma), which is used to assess students' study progress. The dataset holds 10 English sentences and their correct translations into Irish.

Same as for the evaluation campaign in Section 5.1., we performed an automatic evaluation based on the BLUE

[13] These sentences, 400 from each corpus, were not used in any of the training steps when building the translation models. The evaluation data set is available under: http://server1.nlp.insight-centre.org/iris/iris_eval_set.tgz

[14] The order of the added data in the table was determined by the discovery of it during the IRIS development.

[15] Translations done on September 15th 2015

#	Corpus	Evaluation data set [BLEU]	
		English→Irish	Irish→English
0	DGT	9.14	8.74
1	+EU Bookshop	13.83	16.12
2	+EU constitution	12.58	20.98
3	+Focal	12.87	16.69
4	+GNOME	13.52	17.52
5	+Irish legislation	10.48	19.35
6	+KDE4	13.17	20.48
7	+News-2007 (English)	/	25.73
8	+Ubuntu	14.55	25.60
9	+Wikipedia Titles	19.73	28.11
10	+Irish sent. bank	19.90	24.64
11	+Food and Beverages	17.25	25.98
12	+Wikipedia (Irish)	21.09	/
13	+Textbooks	28.55	36.88
14	+Apertium	28.64	31.46

Table 3: Automatic translation evaluation based on BLEU with different parallel/monolingual corpora for the Irish Diploma data set.

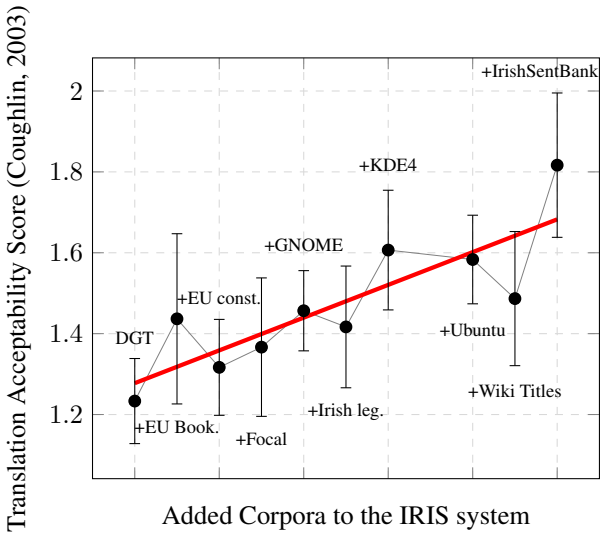

Figure 2: Manual annotation of translation quality (with standard error rate) of Irish sentences based on cumulative training (up to system number 10) of IRIS.

metric. The scores in Table 3 show an improvement in translation quality as more data is added to the IRIS system, although the performance is worse compared to the evaluation set extracted from the used corpora.

Due to the small size of the data set, we asked bilingual evaluators with a background in Irish language teaching to manually inspect the automatically generated translations from English into Irish. For this task, the evaluators annotated sentences produced by IRIS at different stages according to the four acceptability classes defined in Coughlin (2003):

1. Unacceptable: not comprehensible or little information transferred accurately.
2. Possibly Acceptable: possibly comprehensible, some information transferred accurately.
3. Acceptable: Not perfect, but definitely comprehensible, with accurate transfer of all important information.
4. Ideal: Not necessarily a perfect translation, but grammatically correct, with all information accurately transferred.

Similarly to the automatic evaluation, translations generated by early built systems, i.e. systems trained on less knowledge, performed worse compared to translations of systems which learned translation candidates from a larger pool of monolingual or parallel data (Figure 2).

We computed the inter-annotator agreement between human annotators, whereby they achieved an average κ score (Fleiss, 1971) of 0.307, which can be interpreted as fair agreement following Landis and Koch (1977).

Due to the performance difference in terms of BLEU we examined the vocabulary overlap of both evaluation sets and the translation candidates stored the translation models.[16] As illustrated in the upper part of Figure 3, the randomly extracted evaluation set has an evenly distributed vocabulary overlap, with a 90% coverage of uni-grams, 70% of bi-grams and a 20% coverage for five-grams, respectively. We additionally observed that the n-gram overlap between evaluation set and translation models does not grow as more data is added to the system (violet line), which indicates that the translation system learned some of the correct translations within the evaluation set already at an early stage of learning the translation models.

A lower overlap is shown for the Irish diploma evaluation set, where around 70% of uni-grams are only covered, whereby the coverage drops between 20% and 30% for bi-grams. This demonstrates, that this evaluation set is much harder to translate for the IRIS system than the randomly selected evaluation set. Nevertheless, we observed that although the vocabulary overlap (evaluation set vs. translation models) does not grow, the BLEU scores improve as more data is added to the system (purple line). This indicates that the newly added data does not necessary provide new translation candidates regarding to the evaluation set, but can nevertheless help to improve the translation alignment quality of the whole translation model.

6. Conclusion

This paper presented IRIS, a publicly accessible SMT system for translating English into Irish and vice versa. The system provides access to the translation system via a Web service and allows the users to upload monolingual or parallel data. Based on the evaluation data set, IRIS outperformed Google Translate for both translation directions.

[16]The overlap was calculated between source/target n-grams in the evaluation set and the translation candidates within the translation model.

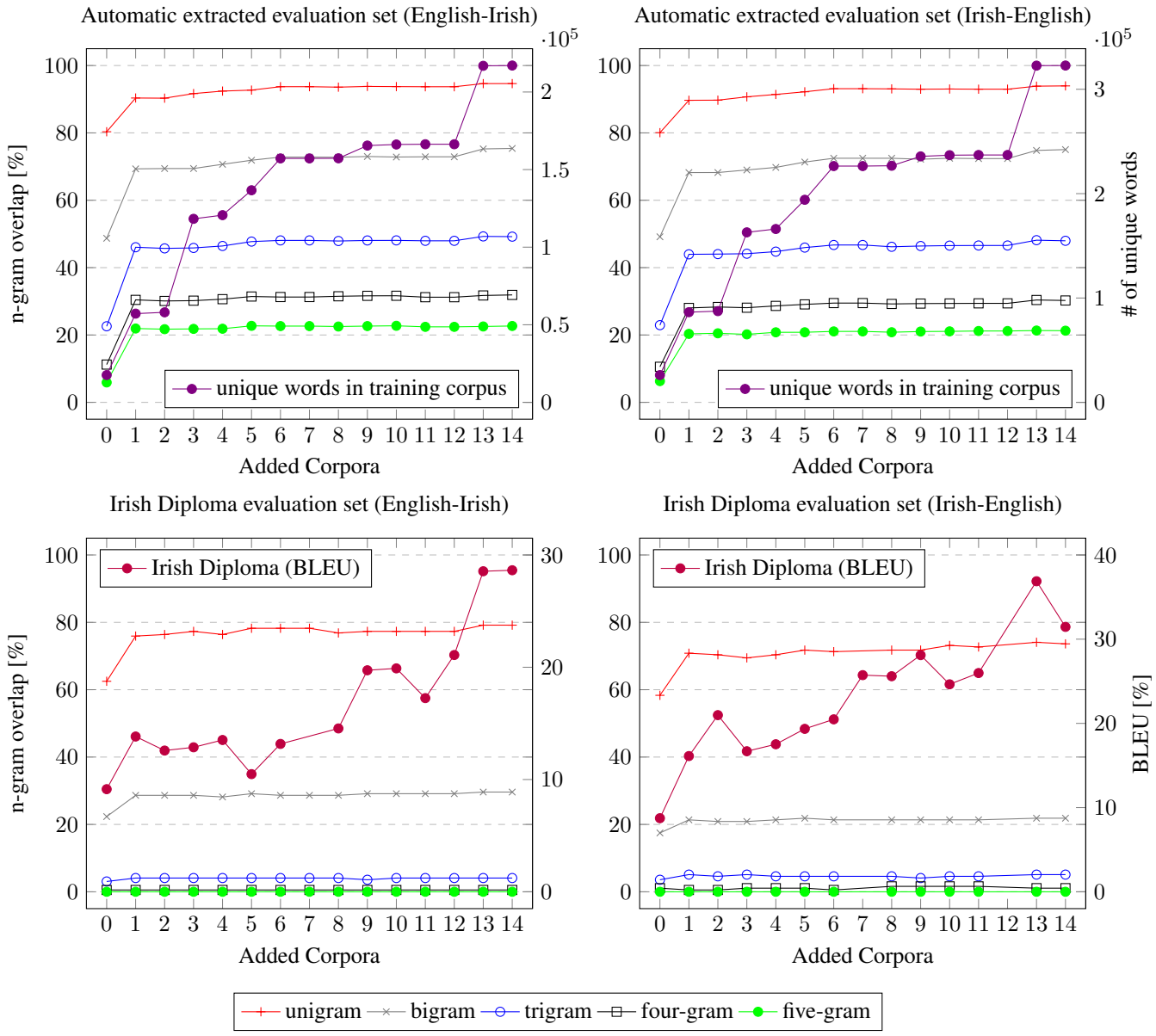

Figure 3: Overlap between n-grams in phrase table compared to the evaluation set in English and Irish

Although the evaluation showed improvement as more data is added to IRIS, the manual evaluators annotated the translation quality rather low. Therefore, we will continue to gather English-Irish monolingual or parallel corpora, which can be embedded into IRIS. Additionally, we will focus on how to better incorporate dictionary/terminological knowledge with sentence aligned corpora for the SMT training.

Acknowledgement

We would like to thank Seathrún Ó Tuairisg, Rosemary Coll and Niall Mac Uidhilin for the manual evaluation of the Irish translations. We would also like to thank the Translation and Interpreting Unit in Acadamh na hOllscolaíochta Gaeilge at the National University of Ireland, Galway for providing us with the additional English-Irish parallel corpus. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight).

7. Bibliographical References

Babych, B., Hartley, A., and Sharoff, S. (2007). Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 412–418, Copenhagen.

Coughlin, D. (2003). Correlating automated and human assessments of machine translation quality. In *Proceedings of the MT Summit IX*.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Dowling, M., Cassidy, L., Maguire, E., Lynn, T., Srivastava, A., and Judge, J. (2015). Tapadóir: Developing a

statistical machine translation engine and associated resources for irish. In *The 4th LRL Workshop: Language Technologies in support of Less-Resourced Languages*.

Fishman, J. A. (1991). *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Multilingual matters. Multilingual matters, Clevedon, United Kingdom.

Fleiss, J. (1971). Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Heafield, K. (2011). KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Judge, J., Ní Chasaide, A., Ní Dhubhda, R., Scannell, K. P., and Uí Dhonnchadha, E. (2012). *An Ghaeilge sa Ré Dhigiteach – The Irish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. pages 127–133, Edmonton, Canada, May/June.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Stroudsburg, PA, USA.

Landis, J. and Koch, G. (1977). Measurement of Observer Agreement for Categorical Data. In *Biometrics*, volume 33.

Lewis, W. D. and Yang, P. (2012). Building mt for a severely under-resourced language: White hmong. Association for Machine Translation in the Americas, October.

Lewis, W. D., Munro, R., and Vogel, S. (2011). Crisis mt: Developing a cookbook for mt in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, July.

Lewis, W. (2010). Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation*. European Association for Machine Translation, May.

Mac Congáil, N. (2004). *Irish Grammar Book*. Cló Iar-Chonnachta, Indreabhán, Conamara, 1 edition.

Ó Broin, B. (2014). New urban irish: Pidgin, creole, or bona fide dialect? the phonetics and morphology of city and gaeltacht speakers systematically compared. *Journal of Celtic Linguistics*, 15(1):69–91, April.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Rannóg an Aistriucháin. (1945). *Litriú na Gaeilge: Lámhleabhar an Chaighdeáin Oifigiúil*. Oifig an tSoláthair, Baile Átha Cliath, Éire.

Rannóg an Aistriucháin. (1958). *Gramadach na Gaeilge agus Litriú na Gaeilge*. Oifig an tSoláthair, Baile Átha Cliath, Éire.

Scannell, K. P. (2006). Machine translation for closely related language pairs. In *Proceedings of the 5th SALTMIL Workshop on Minority Languages and the 5th International Conference on Language Resources and Evaluation (LREC-2006)*.

Scannell, K. P. (2014). Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015). Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translati on (WMT-15)*, pages 256–273, Lisbon, Portugal, September.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey, may.

Tulloch, S. (2006). Preserving dialects of an endangered language. *Current Issues in Language Planning*, 7(2-3):269–286.

Vivian Uíbh Eachach, editor. (2012). *Gramadach na Gaeilge: An Caighdeán Oifigiúil, Athbhreithnithe*. Tithe an Oireachtais, Baile Átha Cliath, Éire.