Developing an interlingual translation lexicon using WordNets and Grammatical Framework

Anonymous

Abstract

Machine translation using the Grammatical Framework (GF) was originally developed for controlled languages with well-defined abstract semantics that acts as an interlingua. More recently, the coverage of GF resource grammars and its processing capabilities have reached to a level, where open-domain tasks such as arbitrary text parsing and translation look a possibility. But, we need several new advances, including robust parsing, parse-tree disambiguation, word sense disambiguation (WSD), and wide-coverage interlingual lexicons. In this paper, we address the later two issues. First, we build a wide coverage interlingual translation lexicon using the Princeton and Universal WordNet data. Then, we propose a technique to do WSD in GF, by integrating an existing WSD tool and replacing the usual GF style lexicons, which give one target word for each source word, by the reported WordNet based lexicons. The result is that, with the help of these new lexicons and WSD, the quality of translations improves in most of the cases, as we show by examples. Both WSD in general, and WordNets are of course well known, but this is the first time these tools have been coupled with GF.

1 Introduction

Interlingual translation has the virtue of easily scaling up to a high number of languages. A contemporary example is Google translate, which deals with all pairs of 60 languages mostly by using English as a pivot language. In this way, it can do with just 2 * 59 = 118 sets of bilingual training data, instead of 60 * 59 = 3540 sets. It would be hard to collect and maintain so many pairs, and in many cases, very little data could be found at all.

The roots of an inter-lingual are much older, perhaps in the medieval idea of a universal grammar (Lyons, 1968). The philosophical argument for an interlingua is that translation means expressing the meaning of the source language expression in the target language. Interlingua then is a universal representation of meaning. This view was developed by the philosopher-mathematicians Descartes and Leibniz. In the recent decades, it has been reflected in the work of (Curry, 1961) where, the interlingua is called tectogrammar, in the Rosetta project (Rosetta, 1994), building on the semantic models of (Montague, 1974), and in the UNL project (Universal Networking Language).

Incidentally, interlingua is also in the heart of modern compiler technology, where for instance the GNU Compiler Collection (Stallman, 2001) uses a shared tree representation to factor out the majority of compilation phases between a high number of source and target languages. Both the scaling up and the shared semantic advantages are exploited here: compiler writers save work, and semantics is preserved by design. A compiler, then, is built as a pipeline with **parsing** from a source language to an **abstract syntax tree**, which is analyzed and optimized in the language-independent phases, and finally **linearized** to a target language. It is easy to see an analogy between this pipeline and the way a human language translator could work. But how to make it real? How to scale up to full the size of natural languages?

In current machine translation research, interlingual methods are marginal, despite the wide practically motivated use of pivot languages in systems like Google translate. Closest to the main stream perhaps is the development of linked WordNets. The original Princeton Wordnet for English (Miller, 1995) defines a set of word senses, which many other wordnets map to other languages. Exact implementations of this idea are Finnish (Lindén and Carlson., 2010) and Hindi (Hindi-WrodNet, 2012).

In the linked Wordnet approach, the Princeton WordNet senses work as an interlingua, albeit only on the level of the lexicon. Carlson and Lindẽn (Lindén and Carlson., 2010) give strong arguments why in fact this is a good way to go, despite the often emphasized fact that different languages divide the world in different ways, so that the senses of their word don't map one to one. The evidence from the English-Finnish case shows that 80% of the mappings are one-to-one and un-problematic. As this part of the lexicon can be easily reused, linguists and system builders can concentrate their effort in the remaining 20%.

The Universal WordNet (de Melo and Weikum, 2009) works on the same lines. Building on the Princeton WordNet, it populates the mappings to over 200 different languages by collecting data from different sources (such as the Wikipedia) and using supervised machine learning techniques to propagate the knowledge and infer more of it. What makes it into a particularly interesting resource is that it is, just like the original Princeton WordNet, freely available under the most liberal licenses.

Grammatical Framework (GF)(Ranta, 2004) is a grammar formalism tool based on Martin

Löf's type theory(Martin-Löf, 1982). From the application point of view, it can be seen as a tool to build interlingua based translation systems. In GF, the translation works in a way analogous to compilers: the source language parser returns an abstract syntax tree, which is then linearized to the target language. The parsing and linearization component are defined by using Parallel Multiple Context-Free Grammars (PMCFG, (Seki et al., 1991), (Ljunglöf, 2004)), which give GF an expressive power between mildly and fully context-sensitive grammars. Thus GF can easily handle with language-specific variations in morphology, word order, and discontinuous constituents, while maintaining a shared abstract syntax.

Historically, the main use of GF has been in controlled language implementations (e.g. (Ranta and Angelov, 2010; Angelov and Enache, 2010; Ranta et al., 2012)) and natural language generation (e.g. (Dymetman et al., 2000)), both applied in multilingual settings with up to 15 parallel languages. In recent years, the coverage of GF grammars and the processing performance has enabled open-domain tasks such as treebank parsing (Angelov, 2011) and hybrid translation of patents (Enache, 2012). The general purpose Resource Grammar Library (RGL)(Ranta, 2011) has grown to 26 languages, reaching from "standard average European" ones to Hindi and Urdu (Prasad and Shafqat, 2012), Japanese, Nepali, Punjabi(Shafqat et al., 2011), Thai, and Chinese.

However, GF's power of interlingual translation has yet not been exploited for arbitrary text parsing and translation. There are a number of challenges in that including robust parsing, parse-tree disambiguation, word sense disambiguation, availability of a wide-coverage interlingual translation lexicon etc. In this paper, first, we shall report an experiment on using the WordNets (i.e. Princeton and Universal) to build an interlingual full-form, multiple senses translation lexicon. Then, we show how these lexicon together with a word sense disambiguation tool can be plugged in a translation pipeline. Finally, we describe an experimental setup and give many examples to highlight the effects of this work.

The paper is structured as follows: Section 2 explains how the Princeton and Universal WordNets are converted to a GF lexicon by using the senses as an abstract syntax and enriching the language mappings with morphological information. Section 3 describes a system architecture and shows where the work reported in this paper fits. Section 4 describes the experimental setup and shows some evaluation results. Section 5 gives a summary of the immediate next steps, based on lessons learned from evaluation results. Section 6 concludes.

2 From Universal Wordnet to a GF Lexicon

As mentioned previously, the original Princeton WordNet(Miller, 1995) defines a set of word senses, and the Universal WordNet(de Melo and Weikum, 2009) maps them to different languages. In this multilingual scenario, the Princeton WordNet senses can be seen as an abstract representation, while the Universal WordNet mappings can be seen as concrete representation of those senses in different languages. GF grammars use very much the same technique of one common abstract and multiple parallel concrete representations to achieve multilingualism. Due to this compatibility, it becomes very natural to easily build a multilingual GF lexicon using data from those two resources (i.e. Princeton and Universal WordNets). This section briefly describes the experiment we did to build one abstract and multiple concrete GF lexicons for a number of languages including German, French, Finnish, Swedish, Hindi, and Bulgarian. The method is very general, so can be used to

build similar lexicon for any other language for which data is available in the Universal WordNet.

2.1 GF Abstract Lexicon

The Princeton WordNet data is distributed in the form of different database files. For each of the four lexical categories (i.e. noun, verb, adjective, and adverb), two files named 'index.pos' and 'data.pos' are provided, where 'pos' is noun, verb, adj and adv. Each of the 'index.pos' files contains all words, including synonyms of the words, found in the corresponding part of speech category. While, each of the 'data.pos' files contains information about unique senses belonging to the corresponding part of speech category. For our purposes, there were two possible choices to build an abstract representation of the lexicon:

1. To include all words of the four lexical categories, and also their synonyms (i.e. to build the lexicon from 'index.pos' files)
2. To include only unique senses of the four categories with one word per sense, but not the synonyms (i.e. to build the lexicon from the 'data.pos' files)

To better understand this difference, consider the words 'brother' and 'buddy'. The word 'brother' has five senses with sense offsets '08111676', '08112052', '08112961', '08112265' and '08111905' in the Princeton WordNet 1.7.1[1], while the word 'buddy' has only one sense with the sense offset '08112961'. Choosing option (1) means that we have to include the following entries in our abstract lexicon.

```
brother_08111676_N
brother_08112052_N
brother_08112961_N
brother_08112265_N
brother_08111905_N
buddy_08112961_N
```

We can see that the sense with the offset '08112961' is duplicated in the lexicon: once with the lemma 'brother' and then with the lemma 'buddy'. However, if we choose option (2), we end up with the following entries.

```
brother_08111676_N
brother_08112052_N
brother_08112265_N
brother_08111905_N
buddy_08112961_N
```

Since the file 'data.noun' lists the unique senses rather than the words, their will be no duplication of the senses. However, the choice has an obvious effect on the lexicon coverage, and depending on whether we want to use it as a parsing or as a linearization lexicon, the choice becomes critical. Currently, we choose option (2) for the following two reasons.

[1] We choose WordNet 1.7.1, because the word sense disambiguator that we are using in our translation pipeline is based on WordNet 1.7.1

1. The Universal WordNet provides mappings for synsets (i.e. unique senses) but not for the individual synonyms of the synsets. If we choose option (1), as mentioned previously, we have to list all synonyms in our abstract representation. But, as translations are available only for synsets, we have to put the same translation against each of the synonym of the synset in our concrete representations. This will not gain anything (as long as we use these lexicon as linearization lexicons), but will increase the size of the lexicon and hence may have a negative impact on the processing speed of the translation system.

2. At the current stage of our experiments we are using these lexicons as linearization lexicons, so having at least one translation of each unique sense is enough.

Our abstract GF lexicon covers 91516 synsets out of around 111,273 synsets in the WordNet 1.7.1. We exclude some of the synsets with multi-word lemmas. We consider them more of a syntactic category rather than a lexical category, and hence deal with them at the syntax level. Here, we give a small segment of our abstract GF lexicon.

```
abstract LinkedDictAbs = Cat ** {

  fun consentaneous_00526696_A : A ;
  fun consecutive_01624944_A  : A ;
  fun consequently_00061939_Adv : Adv ;
  fun abruptly_00060956_Adv  : Adv ;
  fun consequence_09378924_N : N ;
  fun consolidation_00943406_N   : N ;
  fun consent_05596596_N  : N ;
  fun conservation_06171333_N : N ;
  fun conspire_00562077_V : V ;
  fun sing_01362553_V2  : V2 ;
  ........
  ........
 }
```

The first line in the above given code states that the module 'LinkedDictAbs' is an abstract representation (note the keyword 'abstract'). This module extends (achieved by '**' operator) another module labeled 'Cat[2]' which, in this case, has definitions for the morphological categories 'A', 'Adv', 'N' and 'V'. These categories correspond to the 'adjective', 'adverb', 'noun', and 'verb' categories in the WordNet respectively. However, note that in GF resource grammars we have a much fine-grained morphological division for verbs. We sub-categorize them according to their valencies i.e 'V' is for intransitive, and 'V2' for transitive verbs. We refer to (Bringert et al., 2011) for more details on these divisions.

Each entry in this module is of the following general type:

```
fun lemma_senseOffset_t : t ;
```

[2]This module has definitions of different morphological and syntactic categories in the GF resource grammar library

Keyword 'fun' declares each entry as a function of the type 't'. The function name is composed of lemma, sense offset and a type 't', where lemma and sense offset are same as in the Princeton WordNet, while 't' is one of the morphological types in GF resource grammars.

This abstract representation will serve as a pivot for all concrete representations, which are described next.

2.2 GF Concrete Lexicons

We build the concrete representations for different languages using the translations obtained from the Universal WordNet data and GF morphological paradigms (Détrez and Ranta, 2012; Bringert et al., 2011). The Universal WordNet translations are tagged with a sense offset from WordNet 3.0[3] and also with a confidence score. As, an example consider the following segment form the Universal WorNet data, showing German translations for the noun synset with offset '13810818' and lemma 'rest' (in the sense of 'remainder').

```
n13810818 Rest              1.052756
n13810818 Abbrand          0.95462
n13810818 Ruckstand      0.924376
n13810818 Restbetrag      0.662388
n13810818 Restauflage      0.446788
n13810818 Restglied      0.446788
n13810818 Restbestand      0.446788
n13810818 Residuum       0.409192
```

Each entry is of the following general type.

```
posSenseOffset translation confidence-score
```

In cases, where we have more than one candidate translations for the same sense (as in the above case), we select the best one (i.e. with the maximum confidence score) and put it in the concrete grammar. Next, we give a small segment from the German concrete lexicon for the above given abstract lexicon segment.

```
concrete LinkedDictGer of LinkedDictAbs = CatGer ** open
 ParadigmsGer, IrregGer,Prelude in  {

  lin consentaneous_00526696_A =  mkA "einstimmig" ;
  lin consecutive_01624944_A =  mkA "aufeinanderfolgend" ;
  lin consequently_00061939_Adv =  mkAdv "infolgedessen" ;
  lin abruptly_00060956_Adv =  mkAdv "gech" ;
  lin consequence_09378924_N =  mkN "Auswirkung" ;
  lin consolidation_00943406_N =  mkN "Konsolidierung" ;
  lin consent_05596596_N =  mkN "Zustimmung" ;
  lin conservation_06171333_N =  mkN "Konservierung" ;
```

[3]However, in our concrete lexicons we match them to WordNet 1.7.1 for the reasons mentioned previously

```
  lin conspire_00562077_V = mkV "anzetteln" ;
  lin sing_01362553_V2 = mkV2 (mkV "singen" ) ;
  ......
  ......
 }
```

The first line declares 'LinkedDictGer' to be the concrete representation of the previously defined abstract representation (note the keyword 'concrete' at the start of the line). Each entry in this representation is of the following general type:

```
lin lemma_senseOffset_t = paradigmName "translation" ;
```

Keyword 'lin' declares each entry to be a linearization of the corresponding function in the abstract representation. 'paradigmName' is one of the morphological paradigms defined in the 'ParadigmsGer' module. So in the above code, 'mkA', 'mkAdv', 'mkN', 'mkV' and 'mkV2' are the German morphological paradigms[4] for different lexical categories of 'adjective', 'adverb', 'noun', 'intransitive verb', and 'transitive verb' respectively. 'translation' in double quotes is the best possible translation obtained from the Universal WordNet. This translation is passed to a paradigm as a base word, which then builds a full-form inflection table. These tables are then used in the linearization phase of the translation system (see section 4)

Concrete lexicons for all other languages were developed using the same procedure. Table 1 gives some statistics about the coverage of these lexicons.

Language	Number of entries
Abstract	91516
German	49439
French	38261
Finnish	27673
Swedish	23862
Hindi	16654
Bulgarian	12425

Table 1: Lexicon Coverage Statistics

3 System architecture

Figure 1. shows an architecture of the translation pipeline. The architecture is inter-lingual and uses the Resource Grammar Library (RGL) of Grammatical Framework (Ranta, 2011) as the syntax and semantics component, Penn Treebank data for parse-tree disambiguation and IMS(It Makes Sense)(Zhong and Ng, 2010) as a word sense disambiguation tool. Even though the syntax, semantic and parse-tree disambiguation are not the main topics of this paper, we give the full architecture just to show where the work reported in this paper fits. Internal GF resources (e.g. resource grammars and dictionaries) are shown in rectangles

[4]See (Bringert et al., 2011) for more details on these paradigms

while the external components (e.g. PennTreebank and IMS(Zhong and Ng, 2010): a wide coverage word sense disambiguation system for arbitrary text.) are shown in double-stroked rectangles.

The input is parsed using English resource grammar (i.e. EngRG in Figure 1) and a comprehensive English dictionary (i.e. DictEng in Figure 1). In cases where the input is syntactically ambiguous the parser will return more than one parse-trees. These trees are disambiguated using a statistical model build from the PennTreebank data. The best tree is further processed using the input from the IMS to tag the lexical nodes with best sense identifiers. This tree is finally linearized to the target language using the target language resource grammar (i.e. TLRG in Figure 1) together with the target language lexicon (i.e. LinkedDict in Figure 1) discussed in section 2.

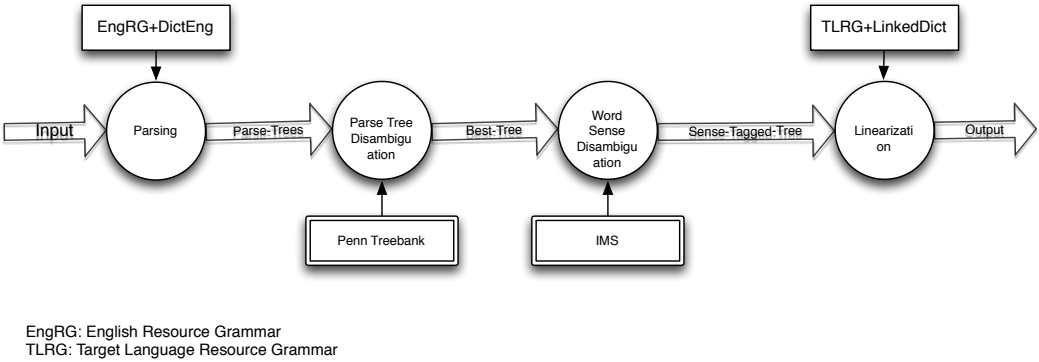

Figure 1: The translation pipeline.

4 Experimental Setup and Evaluation

Our experimental setup is as follows: We take some English text, as source, and translate it to a target language (German and Hindi in these experiments) by passing it through the translation pipeline described in section 3. To show the usefulness of the lexicons described in section 2 and for comparison purposes, we translate the same source twice: with and without word sense disambiguation.

For the first go, we used exactly the same translation pipeline as shown in Figure 1, except that to overcome the deficiencies of our existing parse-tree disambiguator, for some of the examples, we used trees directly from the PennTreebank, which are supposed to be correct. However, this should not damage the claims made in this paper which is about developing wide coverage interlingual translation lexicons and then using them for WSD in an interlingual translation pipeline.

For the second go, we plugged out the word sense disambiguation form the translation pipeline and used our old GF style's (which gives one target word for each source word irrespective of the sense of the source word) lexicons in the linearization phase.

Finally, we compared both candidate translations to find if we have gained anything or not. We did both the manual and automatic evaluations to confirm our findings.

For a set of 25 sentences for English-German pair we got marginal BLEU score improvements (from 0.3904 to 0.399 with 'old' and 'new' dictionaries). Manual inspection, however, was much more encouraging, which also explained the reasons for very low improvements in the BLEU scores in some cases. The reason was this, even if the word sense disambiguation, and hence, our new lexicon gives a better lexical choice, still it will be considered as "wrong" by the evaluation tool, if the gold-standard has a different choice. It was also observed that there were cases, where the 'old' lexicon produced a much better translation than the 'new' one. The reasons for this are obvious. The word sense disambiguator has its own limitations and is known to do mistakes, and also as explained in the section 5 with reasons, the lexicon can not be guaranteed to always give the right translation.

Next, we give a number of example sentence with comments[5] to show that how the new lexicons improved the quality of translations, and also give some examples where it worked the other way around.

4.1 German

1. **Source** He increases the board to seven

   Without WSD `er erhöht das Brett nach einigen sieben`

   With WSD `er vergrößert die Behörde nach einigen sieben`

   Comments `das Brett` is a wooden board (wrong); `erhöht` means "to raise". while `vergrößert` means "increases the size". Note the wrong preposition choice ("to" should be `zu` rather than `nach`). Also, an indefinite determiner (`einige`, some) has been wrongly added to the cardinal number is used as a noun phrase.

2. **Source** the index uses a base of 100 in 1,982

   Without WSD `das Verzeichnis verwendet eine Base nach einige 100 in einigen 1982`

   With WSD `der [index_11688271_N] nutzt einen Operationsbasis von einigen 100 in einigen 1982`

   Comments Note the untranslated word in the WSD version. `Base` means a chemical base, which is the wrong meaning here. `Operationsbasis` is not the obvious choice, but is acceptable.

3. **Source** we were wrong

   Without WSD `wir waren schlecht`

   With WSD `wir waren unkorrekt`

   Comments The translation is not very idiomatic, but `schlecht` simply means "bad", while `unkorrekt` is more specific.

4. **Single words** The choices without and with WSD are shown, followed by comments except if the word was correct. The headword is accompanied by clarifying parenthetical words if needed.

[5]For the comments, we are indebted to Erzsebet Galgoczy and Wolfgang Ahrendt, our colleagues and German informants.

(a) silent (of a film). Without WSD: `schweigsam` usually of a person, WSD: `tonlos`.

(b) (monthly) increase. Without WSD: `Erhöhung`, WSD: `Steigerung` (without context, perhaps the better choice).

(c) set (price for shares) Without WSD: `bereitstellen` (means "provide"), WSD: `bestimmen`.

(d) (item meant for) the future. Without WSD: `die Futur` (means "future tense"), WSD: `die Zukunft`.

(e) (human) adviser. Without WSD: `Ratgeber` (more like a manual), WSD: `Berater` (definitely human).

(f) puzzle. Without WSD: `Puzzle`, WSD: `Rätsel` (more native German).

(g) parts (with liftle context). Without WSD: `Rolle` (refers to a character's part in a play, an unlikely choice of meaning), WSD: `Teile` (parts, more generally).

(h) concerns (in the sense of worries) Without WSD: `Betroffenheiten`, WSD: `Ängste` (not clear which is better).

5. **Source** fear is the father of panic

   With WSD Angst ist der Papa von Angst

   Comment The traditional hilarious machine translation example. It says "fear is fear's daddy".

4.2 Hindi

To represent Hindi, we use an IPA style alphabet, with the usual values and conventions. Retroflexed sounds are written with a dot under the letter: `ṭ, ḍ`, and `ṛ` (a flap) are common, while `ṇ` and `ṣ` occur in Sanskritised Hindi (though many dialects pronounce them `n` and `š`). The palatalised spirant is shown `š` and aspirated stops are shown thus: k^h. A macron over a vowel denotes a long vowel, and `~`, nasalisation. In Hindi, `e` and `o` are always long, so the macron is dropped. Finally, we use `ñ` to mean the nasal homorganic with the following consonant.

Here are examples from our evaluation showing that the WSD system works well; the versions without WSD merely pick the first synonym in the lexicon.

1. **Source** Mr Baris is a lawyer in New York .

   Without WSD `Mr Baris New York mẽ kānūn kā pañḍit hæ`

   With WSD `Mr Baris New York mẽ vakīl hæ`

   Word order `Mr Baris New York in lawyer is`

   Comments `kānūn kā pañḍit` is "expert/teacher in law", while `vakīl` means "lawyer".

2. **Source** we don't depend on pharmaceutical companies for our support

   Without WSD `ham auṣad`^{h}`īya sahyōgī par hamāre b`^{h}`araṇ pōṣaṇ ke liye nahī nirte hæ̃.`

With WSD ham auṣadʰīya kaṁpanī par hamāre nirvāh vyay ke liye nahī ūte hæ̃.

Word order We pharmaceutical companies on our subsistence expenditure for not ??? do

Comments sahyōgī means "company" in the sense of "colleagues", nirvāh vyay means "subsistence expenditure" , while bʰaraṇ pōṣaṇ means "weight bearing". The penultimate word in both versions is nonsense, and the lexicons need to be debugged.

3. **Source** you may recall that a triangle is also a polygon

   Without WSD tum "recall may" ho ki ṭrāyengl "also" bahubʰuj hæ

   With WSD tum smaraṇ kar sakte ho ki trikoṇ bʰī bahubʰuj hæ

   Word order You recall do can that triangle also polygon is

   Comments The version without WSD has several missing words. The WSD version of "recall" is not idiomatic, but understandable.

4. **Single words**
   (a) (human) right. Without WSD: dakṣiṇ right (not left), WSD: adʰikār.
   (b) security (of person). Without WSD: rṇpatr (as in commercial paper), WSD: surakṣā
   (c) property (in law) Without WSD: dʰan (means "wealth"), WSD: sampatti.
   (d) nationality (as citizenship). Without WSD: rāṣṭrīytā (could mean 'nationalism"), WSD: rāṣṭriktā.
   (e) comment. Without WSD: mat prakaṭ (announce opinion), WSD: ṭīkā ṭippaṇī (commentary)
   (f) pair (of socks). Without WSD: pati-patni (married couple), WSD: yugm

   It should be noted that the coverage of the Hindi lexicon is lowest of all the lexicons given in Table 1. The result is that many sentences have missing words in the translations. Further, there is considerable interference with Urdu words (some stemming from the shared base grammar (Prasad and Shafqat, 2012)), and also some mappings coming from the Universal WordNet data are in roman, as opposed to Devanagari (the usual script for Hindi, and what the grammar is based on), so these need to be transcribed. Further, idiomatic phrases are a problem ("before the law" is likely to be rendered "(temporally) before the law" rather than "in the law's eyes").

5 The next steps

Since the Universal WordNet mappings are produced from parallel data by machine learning techniques, the translations can not be guaranteed always for 100% accuracy or for best possible choice. This leaves a window for improvement in the quality of the reported lexicons. One way of improvement is the manual inspection/correction, which may not be an easy task especially for a wide-coverage lexicon with around 100 thousand entries, but is not impossible at the same time. This is once for ever kind of task and will definitely have a strong impact on the quality of the lexicon. Another way is to use manually built WordNets, examples are Finnish, and Hindi WordNets. At the moment, availability of some of

these resources was an issue, so we leave it as a future work. Further, as mentioned in Section 4, Hindi lexicon has some issues related to script, they should be fixed in future.

As mentioned in section 2, GF resource grammars use a much fine-grained morphological categorization for verbs. This division is based on number and type of arguments a verb can take when used as a 'head' in a verb phrase structure. However, this division is not very obvious in some cases. An example is the verb 'admire'. In the current version of the lexicon this verb is tagged as a regular transitive verb (e.g. 'admire_V2'). However, it could have been tagged as 'V2S' as well (i.e. a transitive verb taking a full sentence as an argument). The example usage is in the sentence 'The customers admired the fact that the company has a powerful policy'. Another direction for future work is to explore the verb valencies and enrich the lexicon with such information.

6 Conclusion

We have shown how to use existing lexical resources such as WordNets to develop an interlingual translation lexicon in GF, and how to use it for the WSD task in an arbitrary text translation pipeline. The improvements in the translation quality (lexical), shown by examples in Section 4, are encouraging and is a motivation to work further in this direction. However, it should be mentioned that there is still a lot of work to be done (especially in the open domain text parsing and parse-tree disambiguation phases of the translation pipeline) to bring the translation system to a competitive level.

References

Angelov, K. (2011). *The Mechanics of the Grammatical Framework.* PhD thesis, Chalmers University Of Technology. ISBN 978-91-7385-605-8.

Angelov, K. and Enache, R. (2010). Typeful Ontologies with Direct Multilingual Verbalization. In Fuchs, N. and Rosner, M., editors, *CNL 2010, Controlled Natural Language.*

Bringert, B., Hallgren, T., and Ranta., A. (2011). Gf resource grammar library synopsis. www.grammaticalframework.org/lib/doc/synopsis.html.

Curry, H. B. (1961). Some logical aspects of grammatical structure. In Jakobson, R., editor, *Structure of Language and its Mathematical Aspects: Proceedings of the Twelfth Symposium in Applied Mathematics*, pages 56–68. American Mathematical Society.

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Détrez, G. and Ranta, A. (2012). Smart paradigms and the predictability and complexity of inflectional morphology. In *EACL*, pages 645–653.

Dymetman, M., Lux, V., and Ranta, A. (2000). XML and multilingual document authoring: Convergent trends. In *Proc. Computational Linguistics COLING, Saarbrücken, Germany*, pages 243–249. International Committee on Computational Linguistics.

Enache, Ramona; Espaáa-Bonet, C. R. A. M. L. (2012). A hybrid system for patent translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT12), Trento, Italy.*

Hindi-WrodNet (2012). *Hindi Wordnet. 2012. Universal Word – Hindi Lexicon.* http://www.cfilt.iitb.ac.in.

Lindén, K. and Carlson., L. (2010). Finnwordnet – wordnet pá finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, 17:119–140.

Ljunglöf, P. (2004). *The Expressivity and Complexity of Grammatical Framework.* PhD thesis, Dept. of Computing Science, Chalmers University of Technology and Gothenburg University. `http://www.cs.chalmers.se/~peb/pubs/p04-PhD-thesis.pdf`.

Lyons, J. (1968). Introduction to theoretical linguistics. *Cambridge: Cambridge University Press.*

Martin-Löf, P. (1982). Constructive mathematics and computer programming. In Cohen, Los, Pfeiffer, and Podewski, editors, *Logic, Methodology and Philosophy of Science VI*, pages 153–175. North-Holland, Amsterdam.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Montague, R. (1974). *Formal Philosophy.* Yale University Press, New Haven. Collected papers edited by Richmond Thomason.

Prasad, K. and Shafqat, M. (2012). Computational evidence that hindi and urdu share a grammar but not the lexicon. In *The 3rd Workshop on South and Southeast Asian NLP, COLING*.

Ranta, A. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism. *The Journal of Functional Programming*, 14(2):145–189. `http://www.cse.chalmers.se/~aarne/articles/gf-jfp.pdf`.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Ranta, A. and Angelov, K. (2010). Implementing Controlled Languages in GF. In *Proceedings of CNL-2009, Athens*, volume 5972 of *LNCS*, pages 82–101.

Ranta, A., Détrez, G., and Enache, R. (2012). Controlled language for everyday use: the molto phrasebook. In *CNL 2012: Controlled Natural Language*, volume 7175 of *LNCS/LNAI*.

Rosetta, M. T. (1994). *Compositional Translation*. Kluwer, Dordrecht.

Seki, H., Matsumura, T., Fujii, M., and Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, 88:191–229.

Shafqat, M., Humayoun, M., and Aarne, R. (2011). An open source punjabi resource grammar. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 70–76, Hissar, Bulgaria. RANLP 2011 Organising Committee. http://aclweb.org/anthology/R11-1010.

Stallman, R. (2001). *Using and Porting the GNU Compiler Collection*. Free Software Foundation.

Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics. http://www.aclweb.org/anthology/P10-4014.