Leveraging Morphological Resources in Rules-Based Machine Translation:

The Case for Serbian

Stephen Braich

Department of Computer Science and Engineering
Portland State University
sbraich@pdx.edu

Abstract

The internet provides us with a large corpus of available resources for leveraging morphological resources for machine translation. Today there exists a multitude of tools for morphological analysis to extract lexicons for use in machine translation systems. With the advent of these tools, rules-based machine translation (RBMT), which relies heavily on custom dictionaries, is making a comeback. Here we explore these resources and how we can leverage them building a Mini-Language Resource for the Grammatical Framework (GF)[1].

1. Introduction

In "Outsourcing morphology in Grammatical Framework: a case study for Hungarian" (Listemaa, 2016) advocates for leveraging existing lexicons so that RBMT can focus purely on developing syntax, instead of coding complex morphological rules. We will attempt to replicate the author's work in Serbian drawing on her experience and others who have leveraged resources for other languages.

The goal is to streamline the development of the Grammatical Framework and expand it to as many languages as possible. The Grammatical Framework (GF) (Aarne Ranta, 2014) is a platform for developing multilingual grammar applications. RBMT is a rules-based approach to machine translation that uses a single abstraction grammar called 'interlingua' to translate between languages. If we add a Serbian resource grammar module, GF will be able to translate between all other languages without any further work. This contrasts with neural machine translation[2] (NMT), which requires training a model using a very large corpus of bilingual translations[3].

2. Morphological Resources

We reviewed several lexicons for Serbian, and found the MULTEXT-East Lexicon[4], a monolingual morphologically rich lexicon with an extensive list of surface forms online (Erjavec, 2012). With over

[1] Project Source Code Repository: https://github.com/steve3p0/gf-contrib/tree/master/mini/newmini (forked from https://github.com/GrammaticalFramework/gf-contrib)
[2] Neural Machine Translation (NMT) is the current State of the Art machine translation.
[3] Bilingual corpora requirements for training neural machine translation starts at about 1 million sentences
[4] MULTEXT-East Lexcon: https://www.clarin.si/repository/xmlui/handle/1135/1042

100K lemmas and over 500K surface forms, MULTEXT-EAST needs infinitive and nominative forms mapped to a interlingual lexicon of words in their base form, which I have also found online. Apertium[5], an open source rules based MT platform, has been leveraged by other morphological outsourcing projects (Harsha Vardhan Grandhi, 2015), (Angelov, 2014) (Gregoire Detrez, 2014). It contains a Bosnian/Croatian/Serbian > English RBMT language pair and stores its lexicon as roots without the all the inflected and derivational surface forms.

We were able to successfully replace Grammatical Framework's complex morphological paradigm rules with a simple dictionary lookup that contained all the morphological annotations contained in the MULTI-EAST lexicon. All surface forms entries in the MULTI-EAST lexicon already have POS, tense, person, number, gender, voice and polarity[6]. Once a sentence is parsed in Serbian, the resource grammar in GF, will simply pass the morphosyntactic info to the abstracted interlingual grammar. The interlingual grammar could then automatically transfer and translate into other mini-resource languages. This enables future development of new resource languages to focus on pure syntax (Listemaa, 2016).

Surface Form	Lexeme (infinitive)	Morphosyntactic Annotations
читам	читати	Vm-p1s-an-n---p
читаш	читати	Vm-p2s-an-n---p
чита	читати	Vm-p3s-an-n---p
читамо	читати	Vm-p1p-an-n---p
читате	читати	Vm-p2p-an-n---p
читају	читати	Vm-p3p-an-n---p
књига	књига	Ncfsn--n
књигама	књига	Ncfpd--n
књиге	књига	Ncfpv--n
књиго	књига	Ncfsv--n
књигом	књига	Ncfsi--n
књигу	књига	Ncfsa--n
књизи	књига	Ncfsd--n
књизи	књига	Ncfsl—n

Surface forms for the verb "читати" čitati (to read) and the noun "књига" knjiga (book)

We wrote a Python script that converted the above lexicon into an ordered list of inflections for each person in the present tense. This is in fact how GF already handles exceptions. We simply make the exception the "morphological rule" by expanding out lexicon to contain on surface forms. Below is a two-place verb converted from the MULTI-EAST lexicon. We also added case on the end to tell the mini-resource grammar that the direct object of a 2-place verb must be in the accusative case.

[5] Apertium Bosnian/Serbian/Croatian -> English Resources: https://github.com/apertium/apertium-hbs-eng
[6] Morphosyntactic Tagging of MULTEXT-East: http://nl.ijs.si/ME/V4/msd/html/msd-sr.html

```
--  abstract  func func    P1S     P2S    P3S     P1P      P2P      P3P            Case
lin read_V2 = mkV2 (mkV "читати" "читам" "читаш" "чита" "читамо" "читате" "читају") Acc ;

--  abstract  func Per   Nom-SG   Nom-PL  Acc-SG  Acc-PL
lin book_N  = mkN  Fem  "књига"  "књиге" "књигу" "књиге" ;
```

Grammatical Framework (GF) - Concrete Lexicon in Serbian

3. Grammatical Framework Overview

The Grammatical Framework (GF) uses an abstract syntax as an interlingua and implements languages modules or resources as GF calls them as concrete classes. There are over 30 language resources, but languages of the former Yugoslavia, namely Bosnian/Croatian/Serbian have not been implemented yet. A language resource can be developed as a 'functor' which will make it easier to port it to another language.[7] While this would be ideal if we implement Serbian as a functor language for Bosnian and Croatian, it is out of scope for this project.

3.1 Interlingua – An Abstract Syntax

GF uses an interlingua approach to rules-based machine translation. Instead of having to implement language pairs like English to Serbian and Serbian to English, an intermediate language is 'abstracted' out in the form of a grammar. This abstract syntax must be implemented by all languages resources as 'concrete' syntaxes in the Grammatical Framework. This enables us to implement one language module, Serbian, in this case, and then it can translate from all other resource languages.

As a proof of concept device, GF provides a miniature version of its abstract syntax that we can implement with only a subset of features. Later we can build this out into a full language resource module. Below is a sample of what the no abstract grammar looks like for mini language resources.

```
abstract MiniGrammar = {

  cat
    S ; N ; CN; NP ; V ; VP ; V2 ;
  fun
     UseN      : N   -> CN ;
     UseV      : V   -> VP ;
     PredVP    : NP  -> VP -> S  ;
     ComplV2   : V2  -> NP -> VP ;
     UsePresCl : Pol -> Cl -> S  ;
}
```

[7] Types and Records for Predication, Aarne Ranta, 2014
http://www.aclweb.org/anthology/W14-1401

Here we have defined categories for a sentence, a clause, a noun phrase, a verb phrase, a verb, and a two-place verb. Here are we have defined categories for a sentence, a noun, a common noun, a noun phrase, a verb phrase, a verb, and a two-place verb. And we defined functions for using nouns, verbs, predicate verbs phrases, compliment two-place verb phrases, and for present tense clauses. For brevity, we excluded adverbs, adjectives, prepositions, determiners and many other parts of speech for brevity.

4. Morphology of Concrete Syntax

Next, we will define the Serbian morphological inflection patterns for the parts of speech required by out abstract "interlingua" syntax. As you can see these morphological rules are quite complex and take time to implement. We will compare that a new simplified morphological structure that implements a lookup table using the morphologically annotated lexicon described above.

We implemented the following morphosyntactic features in Serbian:

- Nouns: Person (Singular, Plural)
- Nouns: Gender (Male, Female)
- Nouns: Case (Nominative, Accusative)
- Pronouns: (Nom, Acc)
- Verbs: (Present Tense)
- Verbs: 2-Place Verbs
- Verbs: Polarity: Negative or Positive

For simplicity a lexicon of 100 words is required by the abstract syntax. Aside from the above parts of speech, it also includes adverbs and adjectives. Due to time constraints, morphology for these parts of speech were not implemented and we only implemented the nominative and accusative cases

4.1 Nouns

The mini-language resource lexicon has a collection of 100 nouns, adjectives, and adverbs. The limited size of this lexicon enables the developer to focus on learning the Grammatical Framework. The full language resource modules contain a much larger lexicon. Below is an example of how to implement a concrete morphological syntax for pluralizing Serbian nouns.

```
-- Pluralize Regular Nouns (Nominative Case)
regNoun : Str -> Gender -> Noun = \sg,  g -> mkNoun sg (sg + "и") g;

-- Pluralize Irregular Nouns (Nominative Case)
smartGenNoun : Str -> Gender -> Noun = \stem, g -> case stem of {
  stem1 + v@("а")      => mkNoun  stem  (stem1  +     "е")    g ;
  stem2 + v@("о"|"е")  => mkNoun  stem  (stem2  +     "а")    g ;
  stem3 + v@("к"|"ац") => mkNoun  stem  (stem3  +     "ци")   g ;
  stem4 + v@("да")     => mkNoun  stem  (stem4  +     "дице") g ;
  stem5 + v@("л")      => mkNoun  stem  (stem5  + v + "е")    g ;
  stem6 + v@("д")      => mkNoun  stem  (stem6  + v + "ова")  g ;
  stem7 + v@("з")      => mkNoun  stem  (stem7  + v + "ови")  g ;
  _                    => regNoun stem g -- everything else
  } ;
```

Grammatical Framework (GF) - Concrete Morophological Rule for Nouns – (the old complex way)

We were able to replace the above morphological inflection pattern with simply this rule that maps vocabulary to our lexicon:

```
Noun : Type = { s : genNumStr; c : Case } ;

mkNoun : (_,_,_,_ : Str) -> Case -> Noun
  = \sgn,pln,sga,pla, c -> let s = n.s ! in case c of {
    Acc => table {Sg => sgm ; Pl => plm } ;
    Nom => table {Sg => sgf ; Pl => plf }
    } ;
  } ;
```

Grammatical Framework (GF) - Concrete Morophological Rule for Nouns – “Leveraging MULT-EAST” Lexicon (new way)

The parameters `(_,_,_,_ : Str)` to this function are simply the surface forms in our lexicon for each lexeme:

```
--  abstract  func Per   Nom-SG   Nom-PL  Acc-SG  Acc-PL
lin book_N  = mkN  Fem  "књига"  "књиге" "књигу" "књиге" ;
```

Grammatical Framework (GF) – Concrete Lexicon in Serbian: the noun “књига” knjiga (book)

The above example is actually how the GF implements exceptions or irregular morphology. For lexemes with irregular surface forms, they simply list the full surface forms of the lexeme is a list. Since we have the MULT-EAST lexicon with all surface forms, we can simply convert this lexicon by pivoting it into the format above.

4.2 Verbs

While the Serbian nouns had a consistent pattern for pluralization, Serbian verbs are another story. Below is a list of morphological rules for inflecting Serbian nouns in the present tense. Even with the small subset of verbs in our mini-lexicons there are few examples where these inflection patterns break down.

Employing our “the exception is the rule” pattern for morphology allows us to avoid this problem all together.

```
smartVerb : Str -> Verb = \inf -> case inf of {
-- stem   inf+suffix          1stPSing    2ndPSing   3rdPSing
 stem+v@("мети")      => inf (stem+"мем")(stem+"меш")(stem+"ме")(stem+"мемо")(stem+v+"мете")(stem+"меју");
 stem+v@("ати")       => inf (stem+"ам") (stem+"аш") (stem+"а") (stem+"амо") (stem+v+"ате") (stem+"ају") ;
 stem+v@("ети"|"ити") => inf (stem+"им") (stem+"иш") (stem+"и") (stem+"имо") (stem+v+"ите") (stem+"е");
     _                => inf inf inf inf inf inf inf
     } ;
```

Grammatical Framework (GF) - Concrete Morophological Rule for Verbs – (the old complex way)

Like with nouns, we were able to replace the above morphological inflection pattern for verbs with simply this rule that maps verb and their surfaced forms in our lexicon:

```
    mkV = overload {
      mkV : Str -> Verb = smartVerb ;
      mkV : (_,_,_,_,_,_,_ : Str) -> Verb = mkVerb ;
      } ;
```

Grammatical Framework (GF) - Concrete Morophological Rule for Verbs – “Leveraging MULT-EAST” Lexicon (new way)

Again, the parameters `(_,_,_,_,_,_,_ : Str)` to this function that “make verbs” are simply the surface forms in our lexicon for each verb:

```
  --  abstract  func func    P1S     P2S    P3S     P1P      P2P      P3P            Case
  lin read_V2 = mkV2 (mkV "читати" "читам" "читаш" "чита" "читамо" "читате" "читају") Acc ;
```

Grammatical Framework (GF) - Concrete Lexicon for Serbian Verbs

Again, this is how the GF implements exceptions to the morphological rules to verbs, just as it does for nouns. This allows the developer to focus on syntax and to decouple morphology from tightly coupled morpho-syntactic rules.

5. Evaluation

We do not even attempt to use automated metrics to assess translation quality with metrics like BLEU. This is mini-language resource to which is used in academic settings to teach computational linguistics. There are projects assessing the translation quality on the larger language resource modules. For this project, the metric is time. How much time can be saved implementing a resource grammar leveraging a morphologically annotated lexicon vs. building the morphological rules ourselves?

Below are some estimates that document the hours spent on various modules. Of course, since this was our first grammar that we built, there was an initial steep learning curve, especially in the setup, configuration, testing, and deployment of our grammar. It is important to note that by far the most difficult part was implementing the accusative case when 2-place verbs. This was due to the fact that it was hard to segregate this morpho-syntactic paradigm into separate morphological and syntactic rules.

Module	Subcategory	Morphological Rule Development (old way)	Lexicon Lookup "exception" (new way)
Nouns	Person (Sg, Pl)	16	4
Nouns	Gender (M, F)	8	2
Nouns	Case (Nom, Acc)	24	6
Verbs	Present Tense	8	8
Verbs	2 - Place	8	4
	Totals	64	24
Verbs	Polarity	1	
Pronouns	Case (Nom, Acc)	1	
	Setup	16	
	Testing	24	
	Deployment	16	

Estimate Effort in Hours Developing a Mini-Language Resource for Serbian.

It should be mentioned that if the process of leveraging morphological lexicons is well documented going forward, more time can probably be saved from the total of 24 hours recorded on implementation of the new lookup method.

6. Conclusions

This project clearly showed that leveraging existing morphological resources in the public domain can significantly reduce the amount of time developing a resource grammar for machine translation. It allows developers to focus syntactic rules. At the very least, a heavily morphologically annotated lexicon can be used by developers who are non-native or have very limited proficiency in language resource development.

Further study is required to assess how well this method can be repeated and leveraged by others as well as how well it can scale a mini-resource grammar into a full one. Performance will be a key indicator along with a comparison in development effort. This project also gave the author much needed experience working with rules-based machine translation in the hope of developing hybrid rules plus neural machine methods to hopefully one day bringing machine translation quality on the same of level human translation.

References

Aarne Ranta, K. A. (2014). Large-Scale Hybrid Interlingual Translation in GF: a Project Description. *SLTC 2014 : 5th Swedish Language Technology Conference.* Uppsala, Sweden: University of Gothenburg.

Angelov, K. (2014). Bootstrapping Open-Source English-Bulgarian Computational Dictionary. *Ninth International Conference on Language Resources and Evaluation* (pp. 1018-1023). Reykjavik, Iceland: European Language Resources Association.

Gregoire Detrez, V. M.-C. (2014). Sharing resources between free/open-source rule-based machine translation. *LREC 2014: Ninth International Conference on Language Resources and Evaluation* (pp. 4394-4400). Reykjavik, Iceland: The European Language Resources Association.

Harsha Vardhan Grandhi, S. P. (2015). Automatic conversion of Indian Language Morphological Processors into Grammatical Framework (GF). *Proceedings of the 12th International Conference on Natural Language Processing* (pp. 197-202). Trivandrum, India: Association of Compulational Linguistics.

Listemaa, I. (2016). Outsourcing morphology in Grammatical Framework: a case study for Hungarian. *SLTC 2016: The Sixth Swedish Language Technology Conference.* Umeå, Sweden: Foundations of Language Processing.

Loïc Dugast, J. S. (2007). Statistical Post-Editing on SYSTRAN's Rule-Based Translation System . (p. Proceedings of the Second Workshop on Statistical Machine Translation). Prague, Czech Republic: Association of Computational Linguists.

Nikola Ljubešić, F. K.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. *LREC 2016: Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 4264-4270). Zagreb, Croatia: European Language Resources Association.