Creating inference rules based on GF syntax trees for the FraCaS test suite

Magdalena Siverbo
Centre for Language Technology
Göteborg University

December 2012

1 Introduction

The project described in this document was initiated by Prof. R. Cooper and carried out by the author at the Centre for Language Technology (CLT) at Göteborg University Oct-Dec 2012.

The aim of the project has been to create inference rules using Grammatical Framework (GF) syntax trees for the FraCaS test suite. Several sets of rules have been created, each at a different abstraction level. The idea has been to compare the inference results between these levels, in order to find out how far one can go in abstraction and still get an acceptable degree of correctness. Similar work has previously been done by B. MacCartney and C. Manning [1, 2] and R. Muskens [3].

2 GF trees

The trees used for our inference rules are the syntactic parse trees for the FraCaS test suite, which were created using GF in a previous CLT project beginning in May 2011.[1] In total, there are 346 problem sets (syllogisms) and consequently the same number of inference rules. (All FraCaS syllogisms can be found at `http://www-nlp.stanford.edu/~wcmac/downloads/fracas.xml`.)

Worth mentioning is that some sets are not complete, i.e. in some syllogisms, one or more sentences are either missing or could not be parsed, and are therefore not represented by a tree. There are 11 syllogisms for which this is the case (276, 305, 309, 310, 315, 318, 320, 321, 322, 323 and 324).

[1] `http://projects.grammaticalframework.org/fracas/`

3 Inference rules

In creating inference rules and in testing the FraCaS test suite against them, SWI Prolog has been used. The inference rules are Prolog predicates of the following format:

rule(ID, [Premise$_1$, Premise$_2$, ...], Hypothesis, Question, Answer).

where the premises, hypothesis and question are in the form of GF trees. The rule ID corresponds to the ID of the FraCaS syllogism from which the rule is built. The following is an example of a rule at level 0.

```
rule(r_210_l0,
     ['Sentence'('UseCl'('Present',
                         'PPos',
                         'PredVP'('PredetNP'(all_Predet,
                                             'DetCN'('DetQuant'('IndefArt',
                                                                'NumPl'),
                                                     'UseN'(mouse_N))),
                                  'UseComp'('CompCN'('AdjCN'('PositA'(small_A),
                                                              'UseN'(animal_N))))))),
      'Sentence'('UseCl'('Present',
                         'PPos',
                         'PredVP'('UsePN'(mickey_PN),
                                  'UseComp'('CompCN'('AdjCN'('PositA'(large_A),
                                                              'UseN'(mouse_N)))))))],
     'Sentence'('UseCl'('Present',
                        'PPos',
                        'PredVP'('UsePN'(mickey_PN),
                                 'UseComp'('CompCN'('AdjCN'('PositA'(large_A),
                                                             'UseN'(animal_N))))))),
     'Question'('UseQCl'('Present',
                         'PPos',
                         'QuestCl'('PredVP'('UsePN'(mickey_PN),
                                            'UseComp'('CompCN'('AdjCN'('PositA'(large_A),
                                                                        'UseN'(animal_N)))))))),
     no).
```

It represents the FraCaS syllogism below, where the ID and answer are given at the top (and the hypothesis and question are in reverse order compared to the rule).

```
fracas-210       answer: no
P1 All mice are small animals.
P2 Mickey is a large mouse.
Q  Is Mickey a large animal?
H  Mickey is a large animal.
```

3.1 Levels

The different sets of inference rules are categorized into four levels – level 0 to 3 – and can be described as follows:

Level 0 Basic level; rules are built from the original GF trees.

Level 1 Non-logical terms (in categories Adv, V3, V2, V2V, V, VV, VS, A, Prep and N) substituted with variables.

Level 2 Tense markers substituted with variables.

Level 3 Multi-term constituents without logical content substituted with variables.

3.2 Versions

Apart from the levels mentioned, for each of the levels 1–3 there are also two versions, for which the labels "prudent" and "radical" have been chosen.

Prudent In substituting constants with variables, certain terms are kept as constants, being considered especially significant for the inference in question.

Radical The constants kept for the prudent version (in categories Adv, V3, V2, V2V, V, VV, VS, A, Prep and N) are substituted with variables.

There are some syntactic categories whose members have not been substituted at all, neither in the prudent nor in the radical versions. A few examples of these categories are numerals, personal and relative pronouns, conjunctions and subjunctions. Another example is ellipsis, which has remained marked as such. (See Appendix A for details on what has and has not been substituted.)

At level 0, nothing is substituted, why there is only one version of it.

3.3 Ordered premises

For certain syllogisms in FraCaS, the inference result is clearly dependent on the order of the premises. Thus when applying a rule to a syllogism, a check is performed to make sure that the premises are in the correct order if the syllogism belongs to the group of syllogisms for which the order of premises is significant. If a certain syllogism is not in this group, the order of premises may vary.

3.4 Conditions

From level 1 and up, some (Prolog) conditions are added. The cause for these conditions is that there are cases where certain lexical items in a rule are equivalent but, for syntactic reasons, are represented by different GF categories. One such example is "Nobel prize", which is represented as both `Nobel_prize_N` and `Nobel_prize_N2`, depending on whether there is a modifying expression for the term, like "in literature". For cases like this, a condition is added to account for the correspondence between the items in question.

4 Results

The data that has been used to test the inference rules are the actual FraCaS syntax trees. For a quantitative result, the calculations have been concerned mainly with the number of instances of test syllogisms using a rule other than the one corresponding to it (the one with the same ID number). In the following, we will use the expression "deviation" or "deviating inference" for this kind of phenomenon.

To start with, we will take a look at Table 1, showing the number of "deviations" for each abstraction level and version. It also shows how many of these inferences are correct and how many are incorrect. Please note that the 346 inferences using the rule corresponding to the syllogism (which are all correct) are not taken into account in this table.

	Prudent		**Radical**	
	Correct	**Incorrect**	**Correct**	**Incorrect**
Level 0	0/8	8/8	–	–
Level 1 & 2	5/22	17/22	17/58	41/58
Level 3	39/75	36/75	73/164	91/164

Table 1: Deviating inferences

We see that in moving from one level of abstraction to the next, the effects are noticable, both in the prudent and the radical versions. The number of deviating inferences increases significantly as we move up the abstraction levels. There is also a significant difference between the prudent and the radical version for each of the levels 1–3.[2]

Now, let us go through the inference results for each of the levels in more detail. What we will especially focus on is the degree of inference correctness per syllogism. At each level, in the data collected we can see how many of a syllogism's inferences are correct and how many are incorrect. From this we get a percentage – e.g. if syllogism 264 uses rules 264, 266 and 269 and exactly two of these inferences are correct, we get a correctness percentage of 67% for this syllogism. For the syllogisms with no deviation, i.e. the ones that use only their own rule, the correctness percentage is always 100%.

4.1 Level 0

For level 0, where the inference rules correspond exactly to the original syllogisms, one might expect a one-to-one relationship between rules and test data and thus a 100% inference correctness. However, in the FraCaS test suite, there are a few examples of duplicates, which are there to account for different readings. In other words, a few syllogisms are repeated, with the only difference being the answer which depends on the interpretation. Thus, when testing the

[2] For details on the inference results, see the Excel sheets "FraCaS Inference Diff Data" and "FraCaS Inference Statistics".

FraCaS problem sets against the level 0 rules, certain syllogisms use more than one rule, and since the purpose of the duplicates is to account for different readings, naturally the duplicate rule returns a different result.

In total, there are four pairs of syllogisms using each other's rules, namely 087–088, 129–130, 160–161 and 256–257. (For example, when testing syllogism 087 against the inference rules, both rules 087 and 088 are used, returning two different answers. The same is true for syllogism 088.)

Table 2 shows the correctness percentage for the inferences at level 0. "Without deviation" refers to the syllogims which use only their own rule, while "With deviation" refers to the rest.

	No of syllogisms	**Correctness**
Without deviation	338	100%
With deviation	8	50%

Table 2: Correctness percentages for level 0

4.2 Level 1

As we move up the abstraction levels, we encounter more deviation from the one-to-one relationship between syllogism and rule. At level 0, all deviations return the wrong inference, which is why the degree of correctness is 50% for each syllogism with (one) deviation. At higher levels, however, deviations can still return the correct result. This is where it gets all the more interesting to note not only whether there are deviations, but also whether the result for each deviating inference is correct or not.

Let us look at the different versions of level 1, first the prudent and then the radical, which, being different, of course render different results.

4.2.1 Prudent version

Just as at level 0, the majority of deviating syllogisms in the prudent version of level 1 give 50% in correctness. Table 3 shows the results.

	No of syllogisms	**Correctness**
Without deviation	328	100%
With deviation	2	100%
	3	67%
	12	50%
	1	33%

Table 3: Correctness percentages for level 1 prudent version

4.2.2 Radical version

In the radical version of level 1, there is more deviation and the degrees of correctness is also a bit more evenly distributed compared to the prudent version.

	No of syllogisms	Correctness
Without deviation	300	100%
With deviation	11	100%
	6	67%
	23	50%
	6	33%

Table 4: Correctness percentages for level 1 radical version

4.3 Level 2

The only constants substituted with variables at level 2 are the tense markers. This substitution turns out to have no effect on the inference results. The results for the prudent version of level 2 are therefore equal to the results for the prudent version of level 1, and similarly for the radical versions.

4.4 Level 3

The level 3 substitutions for multi-term constituents gave room for a large number of deviating inferences. The correctness per syllogism has also spread over a larger span compared to levels 1 and 2, not least in the radical version.

4.4.1 Prudent version

In the prudent version of level 3, 300 of the 346 syllogisms are completely without deviation (interestingly and coincidentally exactly as many as in the radical versions of levels 1 and 2). What is fascinating to observe for the syllogisms with deviation is that the the largest “percentage group” is the one with 100% correctness. Also, the number of syllogisms returning results below 50% is marginal.

	No of syllogisms	Correctness
Without deviation	300	100%
With deviation	20	100%
	2	75%
	2	67%
	2	60%
	18	50%
	2	25%

Table 5: Correctness percentages for level 3 prudent version

4.4.2 Radical version

In the results for the radical version of level 3, we can see that there is more variety in correctness percentages. However, the distribution of these is far from even. Table 6 lists the results.

	No of syllogisms	Correctness
Without deviation	268	100%
With deviation	26	100%
	7	75%
	1	67%
	4	60%
	32	50%
	1	44%
	1	40%
	1	38%
	5	25%

Table 6: Correctness percentages for level 3 radical version

5 Summary

Having looked at the results for all the levels and versions, a few general observations can be made.

Firstly, at all levels and versions, the overwhelming majority of syllogisms have no deviation and thus return a result of 100%.

Secondly, among the syllogisms with deviation, there is a clear overall tendency towards a correctness percentage of 100% or 50%. The weight seems to lie either on one of them or divided between the two.

Thirdly, the number of syllogisms rendering results of 1–49% is never higher than the number of those rendering 51–99%. This is an interesting fact worthy to take note of. Let us assume that for each syllogism being tested, one would select the answer returned by most rules applied. For example, if syllogism 264 used three rules returning two "yes" and one "unknown", one would select "yes" as the proposed answer for the syllogism. The difficulty comes in the 50% group. The choice here has to be more arbitrary, for example the answer given by the first rule applied. Following this suggested procedure for the choice of answer, one would get the following, quite encouraging, result (the percentage showing the degree of correctness over the whole test suite).

	Prudent	**Radical**
Level 0	98,8% (342/346)	–
Level 1 & 2	97,4% (337/346)	95,4% (330/346)
Level 3	96,5% (334/346)	93,9% (325/346)

Table 7: Overall correctness for example procedure

This is merely a suggestion for a possible procedure to follow and it is an area where further research could be done. It should also be investigated how results from testing other data than the FraCaS test suite would compare with these results.

References

[1] MacCartney, Bill and Christopher D. Manning. 2008. Modeling Semantic Containment and Exclusion in Natural Language Inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester.

[2] MacCartney, Bill and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the 8th International Conference on Computational Semantics*, pages 140–156, Tilburg.

[3] Muskens, Reinhard. 2010. An Analytic Tableau System for Natural Logic. In Aloni, Bastiaanse et. al. (eds.), *Logic, Language and Meaning*, vol. 6042 of *Lecture Notes in Artificial Intelligence*, pages 104–113, Springer.

Appendices

A Prudent vs Radical

The following elements, in the syllogisms given within brackets, have been kept as constants in the prudent versions while substituted with variables in the radical ones.

```
Adverbs
-------
at_8_am_Adv (284)
at_least_four_times (321)
at_some_time_Adv (316)
at_the_same_time_Adv (317)
by_11_am_Adv (284)
ever_since_Adv (316)
every_month_Adv (307)
every_week_Adv (308)
for_8_years_Adv (319)
for_a_total_of_15_years_or_more_Adv (319)
for_a_year_Adv (298, 301)
for_an_hour_Adv (303)
for_exactly_a_year_Adv (299)
for_more_than_10_years_Adv (319)
for_more_than_two_years_Adv (296)
for_three_days_Adv (315)
for_two_hours_Adv (303, 304)
for_two_years_Adv (295, 298, 299, 301)
friday_13th_Adv (260)
from_1988_to_1992_Adv (330)
here_Adv (126)
in_1990_Adv (330)
in_1991_Adv (277, 278, 279, 280, 281, 282, 283)
in_1992_Adv (258, 277, 278, 279, 280, 281, 282, 283)
in_1993_Adv (252, 253, 254, 255, 256, 257, 326, 327)
in_1994_Adv (307)
in_a_few_weeks_Adv (322)
in_july_1994_Adv (307)
in_march_Adv (316)
in_march_1993_Adv (258)
in_one_hour_Adv (287)
in_the_coming_year_Adv (321)
in_the_past_Adv (316, 321)
in_two_hours_Adv (284, 285, 286, 287, 289, 291)
on_july_4th_1994_Adv (259)
on_july_8th_1994_Adv (259)
on_the_5th_of_may_1995_Adv (314)
on_the_7th_of_may_1995_Adv (314)
over_Adv (259)
```

```
saturday_july_14th_Adv (260)
since_1992_Adv (252, 253, 254, 255, 256, 257)
the_15th_of_may_1995_Adv (314)
too_Adv (149, 155, 156, 175, 176, 178, 179)
twice_Adv (321)
two_years_from_now_Adv (321)
year_1996_Adv (252, 253, 254, 255, 256, 257)
yesterday_Adv (260)


Verbs
-----
continue_V (324)
exist_V (258)
go8travel_V (321)
start_V (259)
stop_V (324)
---
arrive_in_V2 (314)
build_V2 (326, 327)
destroy_V2 (310)
discover_V2 (289)
finish_V2 (326, 327)
found_V2 (258)
get_V2 (025, 041, 057, 073)
have_V2 (115, 118, 122, 131, 132, 136, 138)
last_V2 (259)
lose_V2 (310)
own_V2 (134, 135, 136, 137)
remove_V2 (124, 125)
send_V2 (123)
spend_V2 (285, 286, 289, 291, 292, 293)
take_V2 (136:2&3)
---
bring_V2V (136)
see_V2V (340, 341, 342)
take_V2V (317)
---
rent_from_V3 (131, 132)
tell_about_V3 (318)
---
believe_VS (335)
discover_VS (325)
know_VS (320, 325, 334)
---
can_VV (323)
do_VV (145, 173, 174, 175, 176, 178, 179, 180, 181, 230, 231, 232, 267, 268, 318, 346)
finish_VV (284, 318)
going_to_VV (146)
manage_VV (336)
```

```
need_VV (181)
start_VV (284, 324)
try_VV (337)
want_VV (180)


Adjectives
----------
blue_A (158)
clever_A (217, 219)
competent_A (215)
employed_A (316)
false_A (339)
fast_A (160, 161, 162, 223, 227, 229, 242)
female_A (116)
former_A (198, 199, 200, 201)
large_A (204, 205, 206, 207, 208, 209, 210, 211, 212, 213)
likely_A (101)
many_A (230, 231, 232, 233, 234, 235, 236, 237, 239, 240, 241, 243, 317)
missing_A (124, 125, 126)
own_A (185, 189, 292, 293, 294, 301, 302)
red_A (158, 160, 161, 162)
slow_A (162, 223, 227, 229, 242)
small_A (204, 205, 206, 207, 208, 209, 210, 211, 212, 213)
true_A (338)
unemployed_A (316)


Prepositions
------------
at_Prep (136)
before_Prep (169, 170)
for_Prep (097)
from_Prep (123)
in_Prep (292, 293, 295, 296, 314)
on_Prep (097)
out_of_Prep (124, 125, 126)
part_Prep (014, 046, 062, 078, 137)
possess_Prep (122, 138)
than_Prep (216, 217, 233, 234, 235, 236, 237, 238, 243, 244, 245)
to_Prep (121, 136, 191:2, 192:2, 193:2, 194:2, 195:2)


Nouns
-----
case_N (322)
group_N2 (333)
hour_N (285, 286, 289, 291)
one_N (157, 158, 159, 160, 161, 162, 319)
person_N (091, 092, 093, 218, 325, 333, 344)
```

```
today_N (260, 314)
week_N (317)
year_N (292, 293, 295, 296)
```

The following are never substituted with variables, even in the radical versions.

```
*_AdN
*_AdV
*_Subj
*_Pron
*_Det
*_Card
*_Q*
*_Predet
*_Conj
*_PConj
*_RP
Numerals
```

```
Ellipsis
--------
elliptic_CN (172)
elliptic_NP_Pl (239, 240, 241)
elliptic_NP_Sg (244, 346)
elliptic_V2V (345, 346)
elliptic_VP (145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 172, 175, 176,
178, 179, 180, 181, 191, 192, 193, 194, 195, 230, 231, 232, 245, 267, 268, 318)
elliptic_VPSlash (173, 174)
elliptic_Cl (163)
```