Accuracy evaluation of DELPH-IN analyses

Rebecca Dridan

Universität des Saarlandes

DELPH-IN Summit
Barcelona, July 2009

Outline

Exact match evaluation

Test set	Oracle precision
jh5	0.934
jhpstg-test	0.948
cb	0.817
ws02	0.858

Exact match evaluation

Test set	Oracle precision	Top-1 precision
jh5	0.934	0.656
jhpstg-test	0.948	0.433
cb	0.817	0.204
ws02	0.858	0.388

Exact match evaluation

Test set	Oracle precision	Top-1 precision	MRR
jh5	0.934	0.656	0.742
jhpstg-test	0.948	0.433	0.536
cb	0.817	0.204	0.295
ws02	0.858	0.388	0.470

Exact match evaluation

Test set	Oracle precision	Top-1 precision	MRR
jh5	0.934	0.656	0.742
jhpstg-test	0.948	0.433	0.536
cb	0.817	0.204	0.295
ws02	0.858	0.388	0.470

Exact match asks:

Is the top result the very best analysis the grammar is capable of?

But I want to know:

How good is the top result?

Criteria

granular Ask *How good is the analysis?* rather than *Is it completely correct?*

quantitative Produce a final number to easily measure improvements

descriptive In addition to the number, show where an analysis goes wrong

"semantic" Evaluate predicate argument relations rather than phrase structure

***not* directly comparable to other standards** Don't map to a foreign lossy representation

***but* generally understandable** Allow indirect comparisons to other evaluations, and provide information to an application developer on the suitability of the parser for an application

Proposal

Precision, recall and F-score over triples.

Details

- Based on *ltriples* export format.
  - Important addition compared to *triples* format: **character counts**
- includes ARG relations and features such as TENSE

A side trail turning right leads to Lundadalsbandet and Trulsbu.

_a_q<0:0>	ARG0	_trail_n_1<7:11>
compound<2:11>	ARG1	_trail_n_1<7:11>
compound<2:11>	ARG2	_side_n_1<2:5>
udef_q<2:11>	ARG0	_side_n_1<2:5>
_turn_v_1<13:19>	ARG1	_trail_n_1<7:11>
_right_a_focus<21:25>	ARG1	_turn_v_1<13:19>
_lead_v_to<27:31>	ARG1	_trail_n_1<7:11>
_lead_v_to<27:31>	ARG3	_and_c<52:54>
udef_q<36:63>	ARG0	_and_c<52:54>
proper_q<36:50>	ARG0	named<36:50>(Lundadalsbandet)
_and_c<52:54>	L-INDEX	named<36:50>(Lundadalsbandet)
_and_c<52:54>	R-INDEX	named<56:63>(Trulsbu)
proper_q<56:63>	ARG0	named<56:63>(Trulsbu)
compound<2:11>	SF	prop
compound<2:11>	TENSE	untensed
compound<2:11>	MOOD	indicative
compound<2:11>	PROG	-
compound<2:11>	PERF	-
_side_n_1<2:5>	IND	+
_trail_n_1<7:11>	PERS	3
_trail_n_1<7:11>	NUM	sg
_trail_n_1<7:11>	IND	+
_turn_v_1<13:19>	SF	prop

Basic algorithm

$gold \leftarrow$ list of gold $(rel1, pred, rel2)$ triples
$goldlen \leftarrow$ length of $gold$
$test \leftarrow$ list of parsed $(rel1, pred, rel2)$ triples
$testlen \leftarrow$ length of $test$

for all $(rel1, pred, rel2)$ in $gold$ **do**
    if $(rel1, pred, rel2)$ in $test$ **then**
        $correct \leftarrow correct + 1$
        delete/mark $(rel1, pred, rel2)$ in $test$
    end if
end for

$precision \leftarrow correct/testlen$
$recall \leftarrow correct/goldlen$
$fscore \leftarrow 2 * precision * recall/(precision + recall)$

Basic algorithm

$gold \leftarrow$ list of gold $(rel1, pred, rel2)$ triples
$goldlen \leftarrow$ length of $gold$
$test \leftarrow$ list of parsed $(rel1, pred, rel2)$ triples
$testlen \leftarrow$ length of $test$

for all $(rel1, pred, rel2)$ in $gold$ **do**
    if $(rel1, pred, rel2)$ in $test$ **then**
        $correct \leftarrow correct + 1$
        delete/mark $(rel1, pred, rel2)$ in $test$
    end if
end for

$precision \leftarrow correct/testlen$
$recall \leftarrow correct/goldlen$
$fscore \leftarrow 2 * precision * recall/(precision + recall)$

Refinements

Match text spans rather than predicate names?

This trip goes through forest terrain to the lodge at Ulvaskog, where members of the resistance hid during the war.

>	"Ulvaskog,"	<54:62>	ARG0	"Ulvaskog,"	<54:62>
<	"Ulvaskog, . . . war."	<54:114>	ARG0	"Ulvaskog,"	<54:62>
>	"at"	<51:52>	ARG1	"lodge"	<45:49>
<	"at"	<51:52>	ARG1	"goes"	<10:13>
>	"where . . . the war."	<64:114>	ARG1	"lodge"	<45:49>
<	"where . . . the war."	<64:114>	ARG1	"Ulvaskog,"	<54:62>
>	"through"	<15:21>	PROG	-	
>	"through"	<15:21>	PERF	-	

Precision: 0.961 Recall: 0.937 F-score: 0.949

Refinements

Match text spans rather than predicate names?

>	"Ulvaskog,"	proper_q	<54:62>	ARG0	"Ulvaskog,"	named	<54:62>
<	"Ulvaskog, . . . war."	proper_q	<54:114>	ARG0	"Ulvaskog,"	named	<54:62>
>	"at"	_at_p	<51:52>	ARG1	"lodge"	_lodge_n_1	<45:49>
<	"at"	_at_p	<51:52>	ARG1	"goes"	_go_v_1	<10:13>
>	"where . . . the war."	loc_nonsp	<64:114>	ARG1	"lodge"	_lodge_n_1	<45:49>
<	"where . . . the war."	loc_nonsp	<64:114>	ARG1	"Ulvaskog,"	named	<54:62>
>	"through"	_through_p_dir	<15:21>	ARG1	"goes"	_go_v_1	<10:13>
<	"through"	_through_p	<15:21>	ARG1	"goes"	_go_v_1	<10:13>
>	"through"	_through_p_dir	<15:21>	ARG2	"terrain"	_terrain_n_1	<30:36>
<	"through"	_through_p	<15:21>	ARG2	"terrain"	_terrain_n_1	<30:36>
>	"through"	_through_p_dir	<15:21>	SF	prop		
<	"through"	_through_p	<15:21>	SF	prop		
>	"through"	_through_p_dir	<15:21>	MOOD	indicative		
<	"through"	_through_p	<15:21>	MOOD	indicative		
>	"through"	_through_p_dir	<15:21>	TENSE	untensed		
<	"through"	_through_p	<15:21>	TENSE	untensed		
>	"through"	_through_p_dir	<15:21>	PROG	-		
>	"through"	_through_p_dir	<15:21>	PERF	-		

Refinements

predicate argument relations only?

Count ARGn, R-INDEX,L-INDEX, R-HANDL, L-HANDL, RESTR only.

Closer match to predicate argument evaluations ala CCG or ENJU

	Precision	**Recall**	**F-score**	**Total**	**Average**
OVERALL	0.778	0.775	0.777	45133	61.405
ARG-ONLY	0.720	0.716	0.718	14179	19.291

Table: Granular evaluation of the *jhpstg-test* data set

Refinements

	Precision	Recall	F-score	Total	Average
ARG	0.662	0.645	0.653	228	0.310
ARG0	0.736	0.736	0.736	4094	5.570
ARG1	0.738	0.734	0.736	4835	6.578
ARG2	0.729	0.716	0.722	3298	4.487
ARG3	0.830	0.721	0.772	61	0.083
GEND	0.896	0.869	0.882	327	0.445
IND	0.785	0.782	0.784	2666	3.627
L-HNDL	0.631	0.603	0.616	292	0.397
L-INDEX	0.610	0.594	0.602	515	0.701
MOOD	0.829	0.825	0.827	4497	6.118
NUM	0.793	0.791	0.792	3354	4.563
PERF	0.801	0.792	0.797	3018	4.106
PERS	0.795	0.798	0.797	3635	4.946
PROG	0.801	0.792	0.797	3018	4.106
PRONTYPE	0.874	0.890	0.882	390	0.531
R-HNDL	0.712	0.679	0.695	321	0.437
R-INDEX	0.660	0.647	0.653	535	0.728
SF	0.798	0.793	0.796	5552	7.554
TENSE	0.817	0.813	0.815	4497	6.118

Refinements

	Precision	Recall	F-score	Total	Average
ARG	0.662	0.645	0.653	228	0.310
ARG0	0.736	0.736	0.736	4094	5.570
ARG1	0.738	0.734	0.736	4835	6.578
ARG2	0.729	0.716	0.722	3298	4.487
ARG3	0.830	0.721	0.772	61	0.083
GEND	0.896	0.869	0.882	327	0.445
IND	0.785	0.782	0.784	2666	3.627
L-HNDL	**0.631**	**0.603**	**0.616**	**292**	**0.397**
L-INDEX	**0.610**	**0.594**	**0.602**	**515**	**0.701**
MOOD	0.829	0.825	0.827	4497	6.118
NUM	0.793	0.791	0.792	3354	4.563
PERF	0.801	0.792	0.797	3018	4.106
PERS	0.795	0.798	0.797	3635	4.946
PROG	0.801	0.792	0.797	3018	4.106
PRONTYPE	0.874	0.890	0.882	390	0.531
R-HNDL	**0.712**	**0.679**	**0.695**	**321**	**0.437**
R-INDEX	**0.660**	**0.647**	**0.653**	**535**	**0.728**
SF	0.798	0.793	0.796	5552	7.554
TENSE	0.817	0.813	0.815	4497	6.118

What do the numbers look like?

	Granular evaluation			Exact match evaluation		
Test set	Precision	Recall	F-score	Precision	Recall	F-score
jh5	0.965	0.965	0.965	0.702	0.702	0.702
jhpstg-test	0.919	0.898	0.909	0.456	0.447	0.451
cb	0.891	0.616	0.729	0.250	0.177	0.207
ws02	0.895	0.523	0.660	0.452	0.289	0.352

Table: Granular evaluation and exact match evaluation.

Measured across all items for which a gold standard exists.

What do the numbers look like?

	Sentence accuracy by triples			Exact match evaluation		
Test set	Precision	Recall	F-score	Precision	Recall	F-score
jh5	0.729	0.729	0.729	0.702	0.702	0.702
jhpstg-test	0.480	0.471	0.475	0.456	0.447	0.451
cb	0.271	0.195	0.227	0.250	0.177	0.207
ws02	0.465	0.301	0.365	0.452	0.289	0.352

Table: Sentence accuracy by triples versus exact match accuracy

Measured across all items for which a gold standard exists.

What do the numbers look like?

	All triples			Argument relations only		
Test set	Precision	Recall	F-score	Precision	Recall	F-score
jh5	0.965	0.965	0.965	0.940	0.940	0.940
jhpstg-test	0.919	0.898	0.909	0.870	0.849	0.859
cb	0.891	0.616	0.729	0.813	0.562	0.664
ws02	0.895	0.523	0.660	0.828	0.482	0.609

Table: Granular evaluation of all triples, and argument relations only.

Measured across all items for which a gold standard exists.

Thank You!