A User-oriented Approach to Evaluation and Documentation of a Morphological Analyzer

Gertrud Faaß

NLP Institute, University of Stuttgart
and
Institute for Information Science and Natural Language Processing,
University of Hildesheim

Systems and Frameworks for Computational Morphology
2011

Overview

Background

- "Evaluation" of NLP tools has been an issue for 3 decades
- Taking different perspectives:
  - e.g. ISO / EAGLES[1]
    → focus on *users' perspective*
  - e.g. MorphoChallenge
    → focus on *developers' perspective*

[1] Expert Advisory Group on Linguistic Engineering Standards, 1993-95

ISO 9126/EAGLES EAG-EWG-PR.2

- *Usability* and *Functionality*
- Must contain:
  - Object and user description,
    described as measurable attribute-value pairs
  - Library: test suites and procedures

Current developments, e.g. the MorphoChallenge competition

- Description of the testset (“Goldstandard”) by MorphoChallenge organizers
- Description of the tools (algorithms) and their accuracy by developers
- Each tool’s *Precision* and *Recall* in a summary of the competition given in the proceedings

Terms[2] & their application (1/3)

Verification

The process of ensuring

1. *that the intelligent system conforms to specifications*
2. *its knowledge base is consistent and complete within itself*

Application

- Testing the system according to its design criteria
- Documentation of the criteria, the respective tests, and their outcome:
  - e.g. designing gold analyses (*Accuracy*)
  - Describing parameters that can be set for specific tasks, e.g. fine grained or coarse grained analyses (*Suitability*)
  - Describing input formats (word lists, running text) and possible outputs (*Interoperability*)

[2]Gonzales and Barr (2000), Barr and Klavans(2001)

Terms & their application (2/3)

Validation

The process of ensuring that the output of the intelligent system is equivalent to those of human experts when given the same inputs

Application

Developing guidelines for developers and users
on the expected output for specific language phenomena,
e.g.
how neo-classical forms or loan words shall be analysed
$\longrightarrow$ “stated needs” (no valid-for-all solutions possible)

Terms & their application (3/3)

Evaluation

Serves to describe and to verify to what extent the users' requirements are met by the software

Application

So far (specifically for our task): two kinds of outputs: deep, finely grained analysis of words and lemma+inflectional information only

The Stuttgart Morphological Analyzer (SMOR)[4]

- Finite state morphology,
  runs on the basis of SFST (Schmid (2005))
- Inflection, derivation, compounding
- Lexicon base: IMSLex project
  currently 47,671 German base stems, 528 compounding and 1,691 derivation stems[3], 321 prefixes, 133 suffixes

[3] additional stems are generated automatically
[4] Fitschen (2004), Schmid, Fitschen and Heid (2004)

Aims of the evaluation project[5]

IMS Uni Stgt: Gertrud Faaß, Helmut Schmid, Ulrich Heid

1. Design and full description
   of a significant and reusable testsuite,
   thereby development of guidelines for the morphological analysis of German language phenomena
2. Test of a *frozen* version of SMOR against the testsuite and calculation of accuracies
3. Description of the tool in terms of ISO/EAGLES requirements
4. Publication of the full description for download / online use

[5] Part of D-SPIN (Deutsche Sprachressourcen-Infrastruktur) pre-decessor of CLARIN-D

Methods

Preparation of testsuite and documentation

1. Generate a corpus-derived word list
   and manually describe each of the candidates in terms of its morphological properties
2. Make use of SMOR to get “raw” morphological analyses
   exclude non-words & correct the testsuite manually
3. Describe required input and typical outputs of the tool
   with examples for non-expert users
4. Produce documentation as DocBookXML and html,
   create webpage
   generate all downloadable data as comma separated files (csv)

Testsuite design issues

- Chosing types randomly versus covering the full spectrum of German morphology:

→ Use made of SDEWAC (880 m tokens) to select candidates for the testsuite: random choice
condition: at least medium frequency

- SMOR is known to cover all non-productive (=closed class) words of German

→ Only choose productive word forms, for a start: nouns, verbs and adjectives (1,000 word form types each)

- Thompson (1981):
Negative results are as important as positive ones

→ Do not clean the word list,
but leave non-words to test whether SMOR analyses them

Testsuite creation and calculation of SMOR's accuracy

- Run SMOR with the word form list
  Does SMOR analyse the word form at all?

→ Calculate recall on the level of word forms

- Use SMOR-output (analyses) to generate the testsuite:
  check & correct analyses manually, delete wrong analyses, insert missing analyses

→ Calculate precision on the level of analyses

Results: accuracy

Medium value for nouns,verbs,adjectives:

Recall: 97.96 %
Precision: 84.47 %

(see proceedings for details)

Results: testsuite

Downloadable data (for each of the parts of speech) for 726 nouns, 851 verbs, 854 adjectives

- List of words
- category table
- testsuite

“Demo”

`http://www.ims.uni-stuttgart.de/projekte/dspin/ch02.html`

Results: documentation (in German)

- Description of the tool
  (in-/output formats etc.)
- Description of the testsuite
  (with guidelines for human annotators)
- Description of the validation process and its results

Results: examples

NN-012	Informationsbroschüre	5	in<PREF>format<V>ion<NN><SUFF>Broschüre<+NN><Fem><Nom><Sg>
NN-012	Informationsbroschüre	6	in<PREF>format<V>ion<NN><SUFF>Broschüre<+NN><Fem><Gen><Sg>
NN-012	Informationsbroschüre	7	in<PREF>format<V>ion<NN><SUFF>Broschüre<+NN><Fem><Acc><Sg>
NN-012	Informationsbroschüre	8	in<PREF>format<V>ion<NN><SUFF>Broschüre<+NN><Fem><Dat><Sg>
NN-013	Fachaufsicht	s1	Fach<NN>auf<VPART>sehen<V><SUFF><+NN><Fem><Nom><Sg>
NN-013	Fachaufsicht	s2	Fach<NN>auf<VPART>sehen<V><SUFF><+NN><Fem><Gen><Sg>
NN-013	Fachaufsicht	s3	Fach<NN>auf<VPART>sehen<V><SUFF><+NN><Fem><Acc><Sg>
NN-013	Fachaufsicht	s4	Fach<NN>auf<VPART>sehen<V><SUFF><+NN><Fem><Dat><Sg>

for guidelines, see

`http://www.ims.uni-stuttgart.de/projekte/dspin/ch02s05.html`

for linguistic category description, see

`http://www.ims.uni-stuttgart.de/projekte/dspin/ch03.html`

Open issues

- Linking the documentation with clarin
  virtual language observatory (vlo) `www.clarin.eu/vlo`
- Linking the documentation with SMOR in WebLicht
  `https://weblicht.sfs.uni-tuebingen.de/`
- Offering two levels of granularity
  in SMOR analyses via WebLicht

References I

Bankhardt, C.

D-SPIN - Eine Infrastruktur für Deutsche Sprachressourcen.
Sprachreport, 25(1):30 – 31. 2009.

Baroni, M. and A. Kilgarriff.

Large linguistically-processed web corpora for multiple languages.
In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
Trento, Italy: 87 - 90.
2006.

Barr, V.B. and Klavans, J.L.

Verification and Validation of Language Processing Systems: Is it Evaluation?
ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems.
Toulouse: 34 - 40. 2001.

Belz, A.

That's Nice ... What Can You Do With It?
Computational Linguistics.
Vol. 35(1): 111 - 118. 2009.

Nigel Bevan.

Quality in use: Meeting user needs for quality.
Journal of Systems and Software, 49(1):89 - 96. December 1999.

References II

EAGLES.

Evaluation of Natural Language Processing Systems, EAG-EWG-PR.2.
`http://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html`, accessed 2011-08-22
EAGLES, final report edition. October 1996.

Faaß, G. and U. Heid.

Nachhaltige Dokumentation virtueller Forschungsumgebungen.
Tagungsband: 12. Internationale Symposium der Informationswissenschaft (ISI 2011).
Hildesheim, Germany, 9 - 11th March 2011 to appear.

Faaß, G. and U. Heid and H. Schmid.

Design and application of a Gold Standard for morphological analysis: SMOR in validation.
Proceedings of the seventh LREC conference.
Valetta, Malta, 803 - 810. 2010.

A. Fitschen.

Ein Computerlinguistisches Lexikon als komplexes System [PhD Dissertation].
Volume 10 of *AIMS – Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung*.
Lehrstuhl für Computerlinguistik, Universität Stuttgart, Stuttgart. 2004.

Gonzales, A. and V. Barr.

Validation and verification of intelligent systems - what are they and how are they different?
Journal of Experimental and Theoretical Artificial Intelligence.
12(4). 2000.

References III


Harris, Larry E.
Prospects of Practical Natural Language Systems.
Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics.
Philadephia, Pennsylvania, USA: 129 1980.


R. Hausser, editor.
Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994.
Niemeyer, Tübingen, Germany. 1996.


Hinrichs, M. and T. Zastrow and E. Hinrichs.
WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure.
Proceedings of the 7th international Conference on Language Resources and Evaluation (LREC2010).
Valetta, Malta: 489 - 493 2010.


International Standard ISO/IEC 9126.
Information technology – Software product evaluation – Quality characteristics and guidelines for their use.
International Organization for Standardization, International Electrotechnical Commission, Geneva, Suisse.
1991.


M. King and N. Underwood.
Evaluating symbiotic systems: the challenge.
In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC-2006*,
European Language Resources Association (ELRA)
Genova, Italy: 2475 - 2478 2006.

References IV


Kurimo, M. and M. Varjokallio.
Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – morpho challenge 2008.
In *Working Notes for the CLEF 2008 Workshop*. 2008.


Kurimo, M., S. Virpioja, V.T. Turunen, G.W. Blackwood, and W. Byrne.
Overview and results of Morpho Challenge 2009.
In *Working Notes for the CLEF 2009 Workshop*. 2009.


Lehmann, S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvry, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold.
Tsnlp – Test Suites for Natural Language Processing.
In *Proceedings of the 16th International Conference on Computational Linguistics – Volume 2*.
Copenhagen, Denmark: 711 - 716. 1996.


Mahlow, C. and M. Pietrowski
A Target-Driven Evaluation of Morphological Components for German
In *Searching Answers – Festschrift in Honour of Michael Hess on the Occasion of his 60th birtday*.
Münster: 85 - 99. 2009.


Manzi, S. and M. King and S. Douglas
Working towards User-oriented Evaluation.
Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA 96).
Moncton, New-Brunswick, Canada: 155 - 160. 1996.

References V

Schiller, A.

Deutsche Flexions- und Kompositionsmorphologie mit PC-KIMMO.
In Roland Hausser, editor, *Proceedings, 1. Morpholympics*.
Erlangen, 7./8. March 1994.
Tübingen. Niemeyer. 1996.

Schiller, A. and S. Teufel, C. Stöckert, and C. Thielen.

Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS.
Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen. 1995.

Schmid, H.

A programming language for finite state transducers.
In *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005)*,
Helsinki, Finland. 2005.

Schmid, H. and A. Fitschen, and U. Heid.

A German Computational Morphology Covering Derivation, Composition, and Inflection.
In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*,
European Language Resources Association (ELRA).
Lisbon, Portugal: 1263 - 1266. 2004.

K. Spärck Jones and J. R. Galliers.

Evaluating Natural Language Processing Systems.
Number 1083 in Lecture Notes in Artificial Intelligence. Springer,
Cambridge. 1996.

References VI

Spiegler, S. and C. Monson.

EMMA: A novel Evaluation Metric for Morphological Analysis.
Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).
Beijing, China: 1029 - 1037 2010.

Thompson, B.H.

Evaluation of Natural Language Interfaces to Data Base Systems.
Proceedings of the 19th annual meeting of the Association for Compuational Linguistics (ACL '81).
Stanford, California: 39 - 42. 1981.

Nancy L. Underwood.

Issues in Designing a Flexible Validation Methodology for NLP Lexica.
In A. Rubio, N Gallardo, R. Castro, and A. Tejada, editors:
Proceedings of the First International Conference on Language Resources and Evaluation, volume 1.
Granada, Spain: 129 - 134. 1998.