Commonsense Reasoning Using WordNet and SUMO: a Detailed Analysis

Javier Álvez Itziar Gonzalez-Dios **German Rigau**

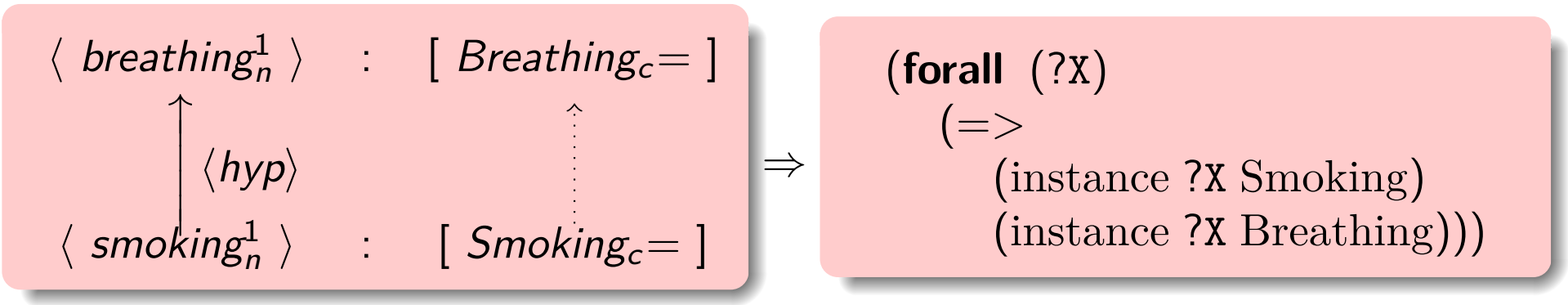
LoRea & **Ixa** Groups
University of the Basque Country (UPV/EHU)

GWC 2019: The 10^{th} Global WordNet Conference
23-27 July, 2019 — Wroclaw, Poland

Outline

Commonsense Workbench & Benchmark

- We describe a detailed analysis of a small sample of a large benchmark of commonsense reasoning problems that have been automatically derived from WordNet, SUMO and their mapping.
- The goal is to assess the quality of both the benchmark and the involved knowledge resources for commonsense reasoning tasks
- Example:



Cross-checking Knowledge Sources

- We exploit the knowledge in the following sources for commonsense reasoning:
  - WordNet (Fellbaum, 1998)
  - SUMO (Niles and Pease, 2001)
  - WN-SUMO Mapping (Niles and Pease, 2003)
- We expect all these knowledge sources to encode correct world knowledge (true knowledge).
- Despite being created manually, they are not free of errors and discrepancies.

Cross-checking Knowledge Sources II

- We apply a new Black-box strategy (Álvez et al., 2019a) to create a large common sense benchmark from these resources.
- The resulting problems are automatically evaluated by means of FOL Automated Theorem Provers (ATPs)
- A detailed analysis is required for a complete assessment:
  - Problems may be solved (yes, no) for bad reasons
    - Expected results do not always indicate a correct ontological modelling
    - Is the knowledge correct in successful tests?
    - Is the knowledge incorrect in failing tests?
  - Problems may remain unsolved (unknown) because of
    - Lack of knowledge in the ontology
    - Lack of resources for ATPs

Outline

SUMO (Niles and Pease, 2001)

- IEEE Standard Upper Ontology Working Group
- SUMO syntax goes beyond first-order logic (FOL)
- SUMO cannot be directly used by FOL Automated Theorem Provers (ATPs) without a suitable transformation
- Different transformations of SUMO into FOL:
  - TPTP-SUMO (Pease and Sutcliffe, 2007)
  - Adimen-SUMO (Álvez et al., 2012)

Mapping

- Mapping between WordNet and SUMO (Niles and Pease, 2003)
- It connects synsets of WordNet to terms of SUMO using 3 relations:
  - *equivalence* (=)
  - *subsumption* (+)
  - *instance* (@)
- Some examples:

$$\begin{aligned} \langle breathing_n^1 \rangle &: [Breathing_c =] \\ \langle education_1^4 \mathrm{n} \rangle &: [EducationalProcess_c +] \\ \langle zero_a^1 \rangle &: [Integer_c @] \end{aligned}$$

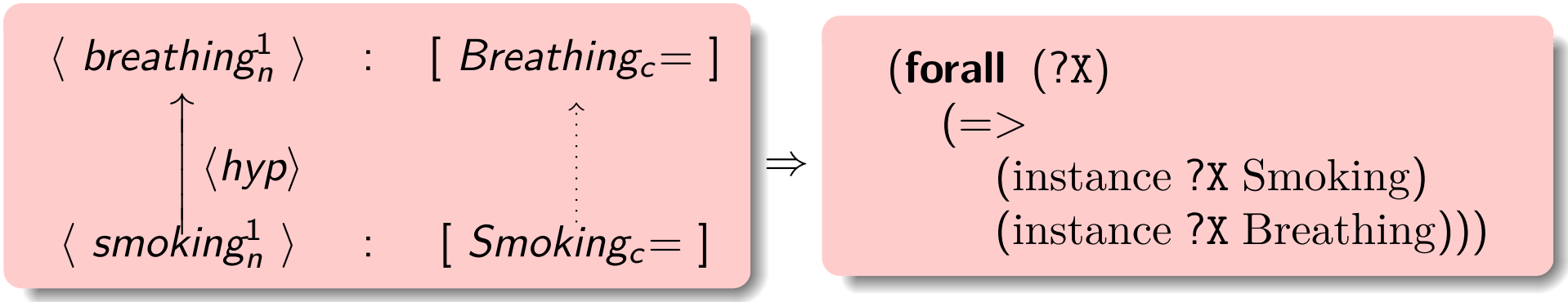
Adimen-SUMO

- Following the line of Horrocks and Voronkov (2006)
- Applying a reengineering process to SUMO (Álvez et al., 2012)
  - With the help of ATPs (*Automated Theorem Provers*)
  - 88 % of the *core* of SUMO (top and middle levels) is translated into FOL
  - Domain ontologies are not used (by now)
- The process of manually debugging the ontology is very costly
  - Only 64 manually created tests were available
  - Example:

```
(forall (?BRAIN ?PLANT)
   (=>
      (and
         (instance ?BRAIN Brain)
         (instance ?PLANT Plant))
      (not
         (properPart ?BRAIN ?PLANT))))
```

Black-box Testing I

- Introduced in Álvez et al. (2015) and fully described in Álvez et al. (2019a)
- Adaptation of the methodology for the design and evaluation of ontologies introduced in Grüninger and Fox (1995)
- Based on the use of Competency Questions (CQs):
  - Problems that an ontology is expected to answer
- CQs are automatically created on the basis of few Question Patterns (QPs) by exploiting WordNet and its mapping into SUMO
- Example:



Black-box Testing II

- Resulting benchmark:

Relation	QP	Problems
Hyponymy	Noun #1	7,539
	Noun #2	1,944
	Verb #1	1,765
	Verb #2	304
Antonymy	#1	91
	#2	574
	#3	2,780
Morphosemantic Links	Agent	829
	Instrument	348
	Result	788
Total	–	**16,972**

Black-box Testing III

- Evaluation is automatic by means of the use of ATPs
- Classification of (dual) problems (truth and falsity tests):
  - *Entailed*: the ATPs are able to demonstrate a truth test
    $\Rightarrow$ Knowledge validation
  - *Incompatible*: the ATPs are able to demonstrate a falsity test
    $\Rightarrow$ Knowledge mismatches due to:
    - WN-SUMO mapping issues
    - WordNet issues
    - SUMO issues
  - *Unresolved*: the ATPs produce no answer within a time limit
    $\Rightarrow$ Missing knowledge ... or insufficient execution time?

Experimental Results

- Using the ATPs Vampire (Kovács and Voronkov, 2013) and E (Schulz, 2002)

QP		#	%	T	E
Noun #1 (7,539)	(+)	3,109	41.24 %	3.92 s.	472.51
	(−)	1,736	23.03 %	53.60 s.	71.43
Noun #2 (1,944)	(+)	1,222	62.86 %	3.82 s.	1,261.07
	(−)	198	10.19 %	132.92 s.	65.75
Verb #1 (1,765)	(+)	587	33.26 %	4.20 s.	391.96
	(−)	260	14.73 %	60.27 s.	54.12
Verb #2 (304)	(+)	137	45.07 %	4.41 s.	1,300.31
	(−)	16	5.26 %	141.73 s.	19.83
Antonym #1 (91)	(+)	29	31.87 %	22.82 s.	419.97
	(−)	4	4.40 %	3.26 s.	433.03
Antonym #2 (584)	(+)	161	27.57 %	116.33 s.	40.95
	(−)	25	4.28 %	0.84 s.	1,410.65
Antonym #3 (2,780)	(+)	978	35.18 %	180.78 s.	45.80
	(−)	9	0.32 %	55.70 s.	17.98
Agent (829)	(+)	39	4.70 %	6.28 s.	0.49
	(−)	3	0.36 %	402.85 s.	0.03
Instrument (348)	(+)	611	17.53 %	45.61 s.	0.23
	(−)	1	0.29 %	595.03 s.	0.00
Result (788)	(+)	94	11.93 %	11.04 s.	0.29
	(−)	11	1.42 %	186.29 s.	0.28
Total (16.972)	**(+)**	6,967	41.05 %	35.20 s.	459.79
	(−)	2,263	13.33 %	62.61 s.	83.33

Outline

Detailed Analysis

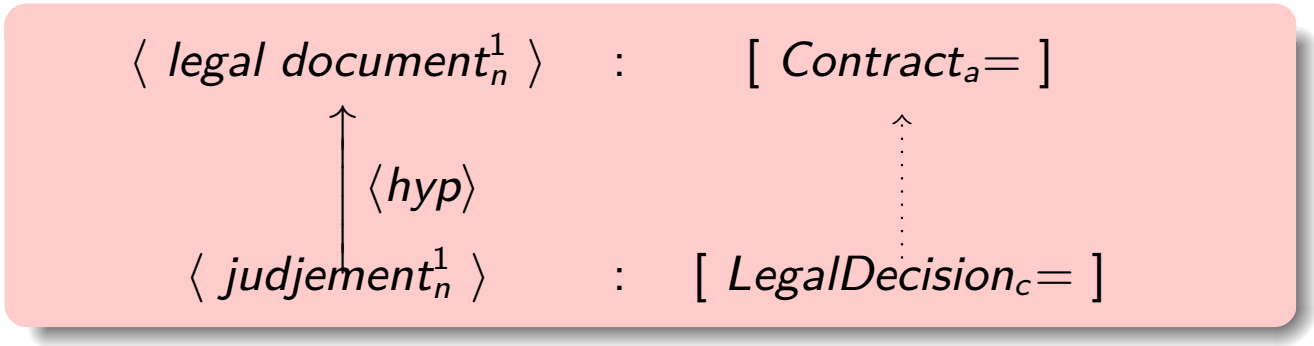
- Randomly selected a sample of 169 problems (1 %)
- We manually inspect:
  - Quality of the mapping of the synset pair
    - Correct & Precise (C&I)
    - Only correct (C)
    - Incorrect (I)
  - Knowledge required for solving a problem
    - Correct (C)
    - Incorrect (I)

Detailed Analysis: Entailed Problems I

QP	#	EP	Mapping			Knowledge	
			C&P	C	I	C	I
Noun #1 (7,539)	80	39	5	28	6	39	0
Noun #2 (1,944)	15	9	5	4	0	9	0
Verb #1 (1,765)	13	5	1	2	2	5	0
Verb #2 (304)	2	0	0	0	0	0	0
Antonym #1 (91)	1	0	0	0	0	0	0
Antonym #2 (584)	6	1	0	0	1	1	0
Antonym #3 (2,78)	33	9	0	4	5	9	0
Agent (829)	5	1	0	1	0	1	0
Instrument (348)	2	0	0	0	0	0	0
Result (788)	12	1	0	1	0	1	0
Total problems (16,972)	**169**	**65**	**11**	**40**	**14**	**65**	**0**

Detailed Analysis: Entailed Problems II

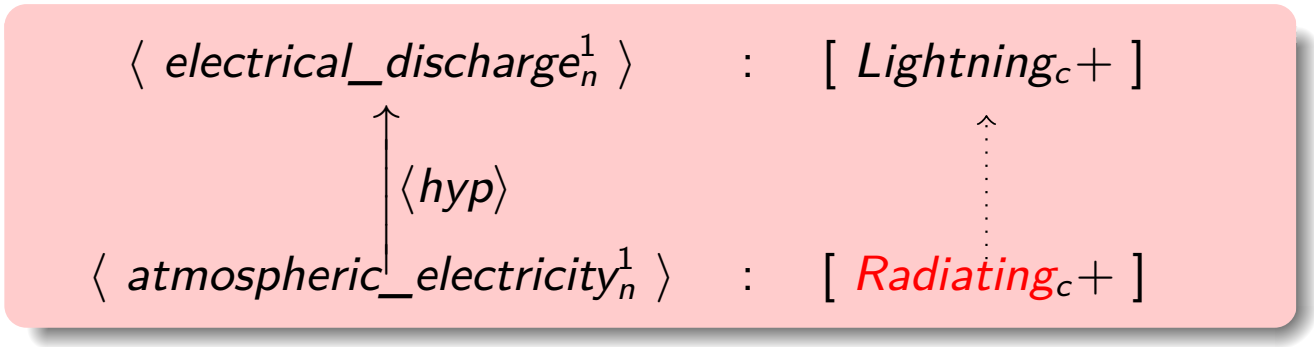
- Case 1: Correct mapping
  - Example:



  - Ontology and mapping knowledge is well-aligned
  - 51 problems (78 % of entailed problems)

Detailed Analysis: Entailed Problems III

- Case 2: Incorrect mapping
  - Example:



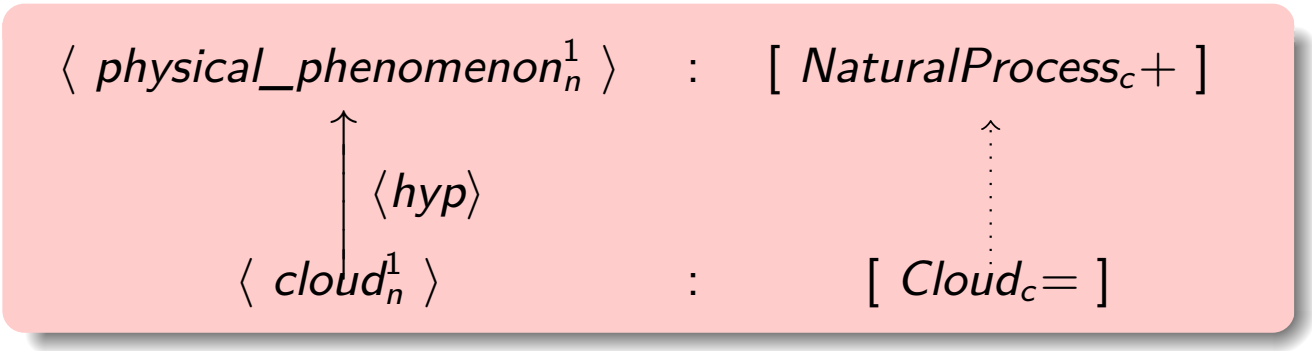
  - Resolved by chance:
    - *Radiating$_c$*: *"Processes in which some form of electromagnetic radiation e.g. radio waves, light waves, electrical energy, etc. is given off or absorbed by something else"*
  - 14 problems (22 % of entailed problems)

Detailed Analysis: Incompatible Problems I

QP	#	IP	Mapping			Knowledge	
			C&P	C	I	C	I
Noun #1 (7,539)	80	15	0	7	8	15	0
Noun #2 (1,944)	15	2	2	0	0	2	0
Verb #1 (1,765)	13	0	0	0	0	0	0
Verb #2 (304)	2	0	0	0	0	0	0
Antonym #1 (91)	1	0	0	0	0	0	0
Antonym #2 (584)	6	0	0	0	0	0	0
Antonym #3 (2,78)	33	0	0	0	0	0	0
Agent (829)	5	0	0	0	0	0	0
Instrument (348)	2	0	0	0	0	0	0
Result (788)	12	0	0	0	0	0	0
Total problems (16,972)	**169**	**17**	**2**	**7**	**8**	**17**	**0**

Detailed Analysis: Incompatible Problems II

- Case 1: Knowledge misalignment
  - Example:



  - *Cloud*$_c$ is subclass of *Substance*$_c$ and *NaturalProcess*$_c$ is subclass of *Process*$_c$, which are disjoint in SUMO
  - 9 problems with *Correct&Precise* or only *Correct* mapping (53 % of incompatible problems)

Detailed Analysis: Incompatible Problems III

- Case 2: Imprecise (*not equivalent*) mapping
  - Example:

$$[\ \mathit{Transfer}_c{=}\ ]\ :\ \langle\ \mathit{fetch}_v^1\ \rangle\ \not\leftrightarrow\ \langle\ \mathit{carry\_away}_v^1\ \rangle\ :\ [\ \mathit{Removing}_c{+}\ ]$$

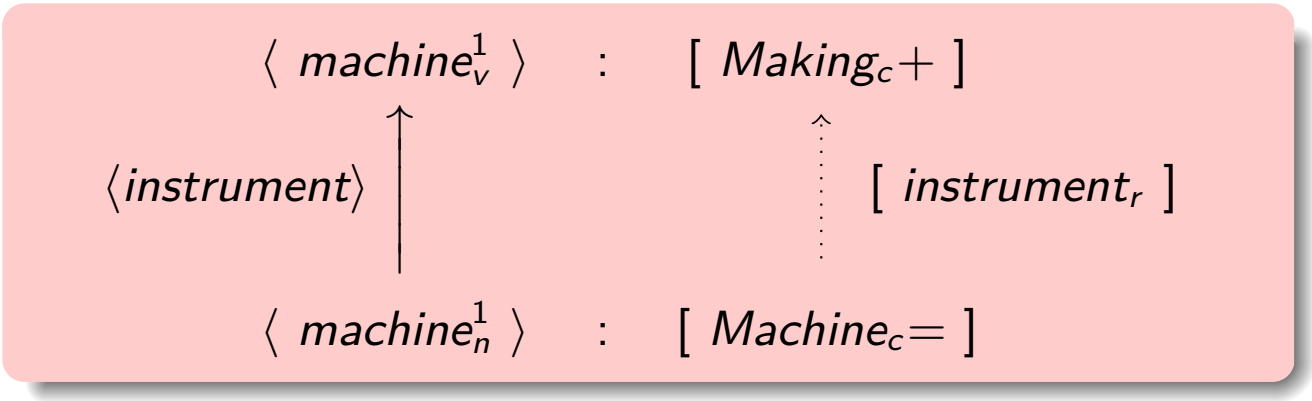
  - $\mathit{Removing}_c$ is subclass of $\mathit{Transfer}_c$ in SUMO
  - The mapping of fetch_v^1 to $\mathit{Transfer}_c$, although correct, is too general
  - 7 problems (41 % of incompatible problems)

Detailed Analysis: Unresolved Problems I

QP	#	UP	Mapping		
			C&P	C	I
Noun #1 (7,539)	80	26	0	19	7
Noun #2 (1,944)	15	4	2	1	1
Verb #1 (1,765)	13	8	0	6	2
Verb #2 (304)	2	2	1	1	0
Antonym #1 (91)	1	1	0	0	1
Antonym #2 (584)	6	5	1	2	2
Antonym #3 (2,78)	33	24	0	7	17
Agent (829)	5	4	1	2	1
Instrument (348)	2	2	2	0	0
Result (788)	12	11	4	2	5
Total problems (16,972)	**169**	**87**	**11**	**40**	**36**

Detailed Analysis: Unresolved Problems II

- Case 1: Lack of knowledge
  - Example:



  - *Machine*$_c$ and *Making*$_c$ are not related in SUMO
  - 45 problems (52 % of unresolved problems)

Detailed Analysis: Unresolved Problems III

- Case 2: Lack of resources
  - Example:

$$[\ Male_a+\ ] \quad : \quad \langle\ male_a^3\ \rangle \quad \not\leftrightarrow \quad \langle\ female_a^1\ \rangle \quad : \quad [\ Female_a=\ ]$$

  - Although it is inferred from SUMO, ATPs cannot find a proof within the given resources
  - 6 problems (7 % of unresolved problems)

Outline

Conclusions and Future Work

- Although 54 % of the problems are solved, only 36 % are resolved for the good reasons:
- The mapping requires a general revision and correction
  - In particular, the mapping of adjectives
- The knowledge in SUMO seems to be correct, but insufficient
- Incompatible problems enable the detection of misalignments between WordNet and SUMO
- Unresolved problems can be used to augment SUMO
- Some problems cannot be resolved because of limitations of ATPs

Bibliography I

J. Álvez, P. Lucio, and G. Rigau. Adimen-SUMO: Reengineering an ontology for first-order reasoning. *Int. J. Semantic Web Inf. Syst.*, 8(4):80–116, 2012.

J. Álvez, P. Lucio, and G. Rigau. Improving the competency of first-order ontologies. In K. Barker and J. M. Gómez-Pérez, editors, *Proc. of the 8th Int. Conf. on Knowledge Capture (K-CAP 2015)*, pages 15:1–15:8. ACM, 2015. ISBN 978-1-4503-3849-3. doi: 10.1145/2815833.2815841.

J. Álvez, P. Lucio, and G. Rigau. A framework for the evaluation of SUMO-based ontologies using WordNet. *IEEE Access*, 7:36075–36093, 2019a. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2904835.

C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

M. Grüninger and M. S. Fox. Methodology for the design and evaluation of ontologies. In *Proc. of the Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI 1995)*, 1995.

I. Horrocks and A. Voronkov. Reasoning support for expressive ontology languages using a theorem prover. In J. Dix et al., editor, *Foundations of Information and Knowledge Systems*, LNCS 3861, pages 201–218. Springer, 2006.

Bibliography II

L. Kovács and A. Voronkov. First-order theorem proving and Vampire. In N. Sharygina and H. Veith, editors, *Computer Aided Verification*, LNCS 8044, pages 1–35. Springer, 2013. ISBN 978-3-642-39798-1.

I. Niles and A. Pease. Towards a standard upper ontology. In Guarino N. et al., editor, *Proc. of the 2nd Int. Conf. on Formal Ontology in Information Systems (FOIS 2001)*, pages 2–9. ACM, 2001. doi: 10.1145/505168.505170.

I. Niles and A. Pease. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In H. R. Arabnia, editor, *Proc. of the IEEE Int. Conf. on Inf. and Knowledge Engin. (IKE 2003)*, volume 2, pages 412–416. CSREA Press, 2003. ISBN 1-932415-08-4.

A. Pease and G. Sutcliffe. First-order reasoning on a large ontology. In Sutcliffe G. et al., editor, *Proc. of the Workshop on Empirically Successful Automated Reasoning in Large Theories (CADE-21)*, CEUR Workshop Proceedings 257. CEUR-WS.org, 2007.

S. Schulz. E - A brainiac theorem prover. *AI Communications*, 15(2-3):111–126, 2002. ISSN 0921-7126.