

Edinburgh, April 17, 2007



Hybrid machine translation: Combining rule-based and statistical MT systems

Andreas Eisele

Saarland University & DFKI, LT Lab

Joint work with Kamran Azam, Yu Chen,
Christian Federmann, Greg Gulrajani,
Eva Hasler, Michael Jellinghaus, Robert
Neßelrath, Armin Schmidt, Silke Theison

e-mail: eisele@dfki.de
WWW: <http://www.dfki.de/lt>

Structure of presentation



- Complementary advantages of rule-based and statistical MT
- Using a SMT decoder to merge outputs of multiple MT engines
- Feeding SMT lexicons into rule-based MT engines
- Thoughts on deeper integration



EuroMatrix: situation in early 2005



MT systems per language pair (data taken from J.Hutchins' Compendium of Translation Software, 12th Edition)

	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Poli.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English		47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48		24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40	23		11	13	8	4	1	1	3	1	-	-	-	-	-	-	-	-	-	-	-	-
Spanish	41	7	11		9	8	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Italian	29	10	13	9		4	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Portuguese	29	5	7	8	4		1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dutch	10	2	4	1	1	1		-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Polish	7	2	1	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Latvian	2	1	1	1	1	1	1	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-
Greek	3	-	3	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-
Czech	1	1	1	-	1	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-
Hungarian	2	2	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-
Swedish	2	1	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-
Finnish	2	1	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-
Slovak	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
Romanian	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-
Danish	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Most language pairs remain uncovered

EuroMatrix: situation in early 2005



MT systems per language pair (data taken from J.Hutchins' Compendium of Translation Software, 12th Edition)

	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Poli.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English		47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48		24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40																						
Spanish	41																						
Italian	29																						
Portuguese	29																						
Dutch	10																						
Polish	7																						
Latvian	2																						
Greek	3																						
Czech	1																						
Hungarian	2																						
Swedish	2																						
Finnish	2																						
Slovak	-																						
Romanian	1																						
Danish	-	1																					
Bulgarian	-																						
Slovene	-																						
Maltese	-																						
Lithuanian	-																						
Irish	-																						
Estonian	-																						

Most language pairs remain uncovered, but some are densely populated

Rule-based vs. statistical MT



Some examples (translate pro \leftrightarrow SMT Koehn 2005)

EN: I wish the negotiators continued success with their work in this important area .

DE: Ich wünsche den Unterhändlern auch weiterhin viel Erfolg auf diesem wichtigen Gebiet.

RBMT: Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich fortsetzen.

continued: verb instead of adjective

SMT: Ich wünsche der Verhandlungsführer fortgesetzte Erfolg bei ihrer Arbeit in diesem wichtigen Bereich.

three inflectional endings wrong

Rule-based vs. statistical MT



More examples

EN: We seem sometimes to have lost sight of this fact .

RBMT: Wir scheinen manchmal Anblick dieser Tatsache verloren zu haben.

idiomatic expression not known

SMT: Manchmal scheinen wir aus den Augen verloren haben, diese Tatsache.

wrong attachment of „diese Tatsache“

Rule-based vs. statistical MT



More examples

EN: I would like to close with a procedural motion .

DE: Mit einem Antrag zur Geschäftsordnung komme ich zum Schluss .

RBMT: Ich möchte mit einer verfahrenstechnischen Bewegung schließen.

wrong translation of procedural motion

SMT: Ich möchte abschließend eine Frage zur Geschäftsordnung.

verb is missing

Rule-based vs. statistical MT



More examples

EN: The leaders of Europe have not formulated a clear vision .

DE: Die Führung Europas hat keine klaren Visionen formuliert .

RBMT: Die Leiter von Europa haben keine klare Vision formuliert.

„Leiter von Europa“ sounds very odd

SMT: Die Führung Europas nicht formuliert eine klare Vision .

syntactically illformed

Motivation of Hybrid MT Approaches



In the early 90s, statistical and rule-based approaches were seen in strict contrast. But PROs and CONs are complementary:

	Syntax	Structural Semantics	Lexical Semantics	Lexical Adaptivity
Rule-based MT	++	+	-	--
Statistical MT	--	--	+	+
Example-based MT	-	--	-	++

→ It is now more or less consensus to target integrated approaches

Two Different Types of Hybridisation



- **Deep Integration:** Design a new setup that combines the advantages of two paradigms, e.g. by integrating some good features of Approach B into Approach A, such as
 - Making a rule-based system adaptive by adding a module for rule learning
 - Making a SMT system syntax-aware by adding syntactical constraints/rules

- **Shallow Integration:** Integrate two or more systems, following different approaches, into a larger system

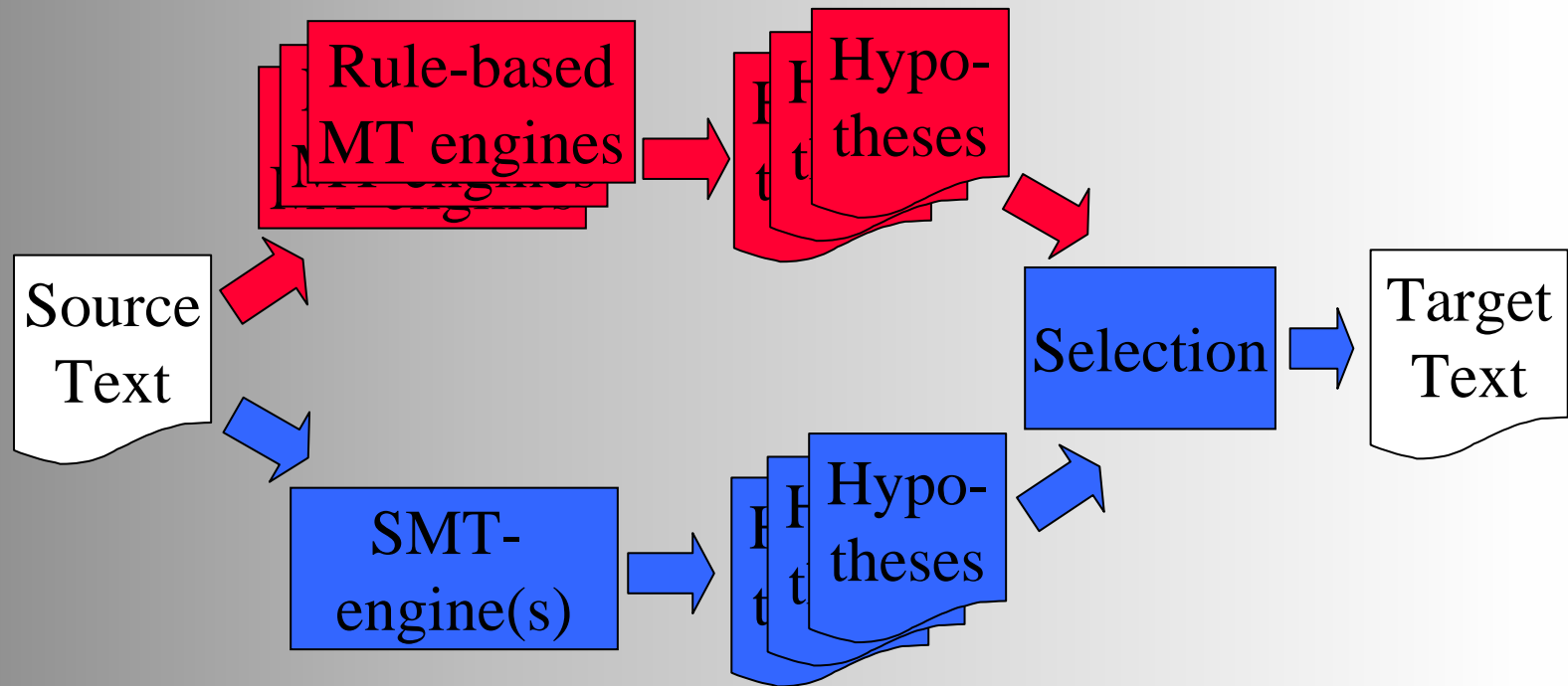
Deep integration is superior in the long run, but also much more challenging. WP2 (Richer models for statistical translation, U Edinburgh) and WP3 (Tree-Based Transfer Models, Charles U Prague) are steps towards deep integration

WP6 (Saarland U) will pursue the modest goal of shallow integration into a MEMT architecture, hopefully still giving insights into the relevant issues related to deep integration

Hybrid MT Architecture I



*Multi-engine MT via black-box integration
(as done in VerbMobil and earlier)*

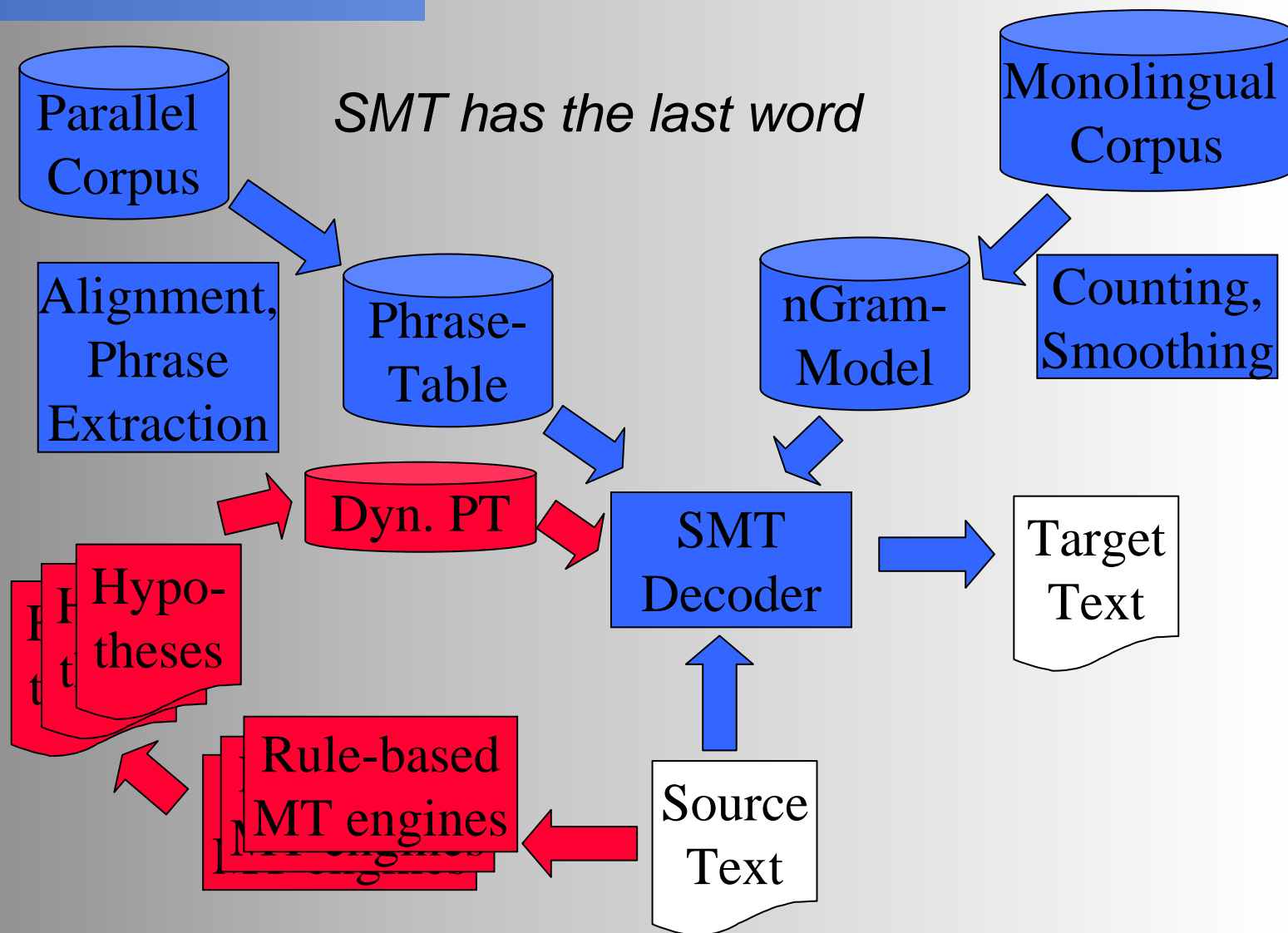


Towards better hybrid MT Architectures



- Disadvantage of simple selection:
For longer sentences, no result will be perfect; we want to combine better parts of multiple outcomes.
- But recombination can be fairly complex, as corresponding parts of alternative candidates are not obvious
- We need
 - alignment of MT results (needs to cope with MT errors)
 - search for best combination
- We can use existing SW modules for both purposes in first steps, error analysis may then suggest improvements

Hybrid MT Architecture II



Hybrid MT Architecture II



Key features:

- Source text is sent through many MT engines, including web-based and locally installed ones
- Alignment between MT output and source text is done via (modified) GIZA++
- Alignment quality is improved by using models trained on larger datasets
- Alignment waiting times are reduced by client-server setup

Hybrid MT Architecture II



Current status:

- Using 6..7 engines: Systran, SDL, ProMT, OpenLogos, translate pro, L&H PTP, Lucy
- Each of the MT engines has its own peculiarities that require attention (encoding, tokenisation, ...)
- Combined phrasables slow down decoding, makes MERT more difficult
- Delays due to technical problems while preparing WMT07 submission
- Implementation essentially done, but current configuration does not yet beat baseline

Hybrid MT Architecture II



Indicative results

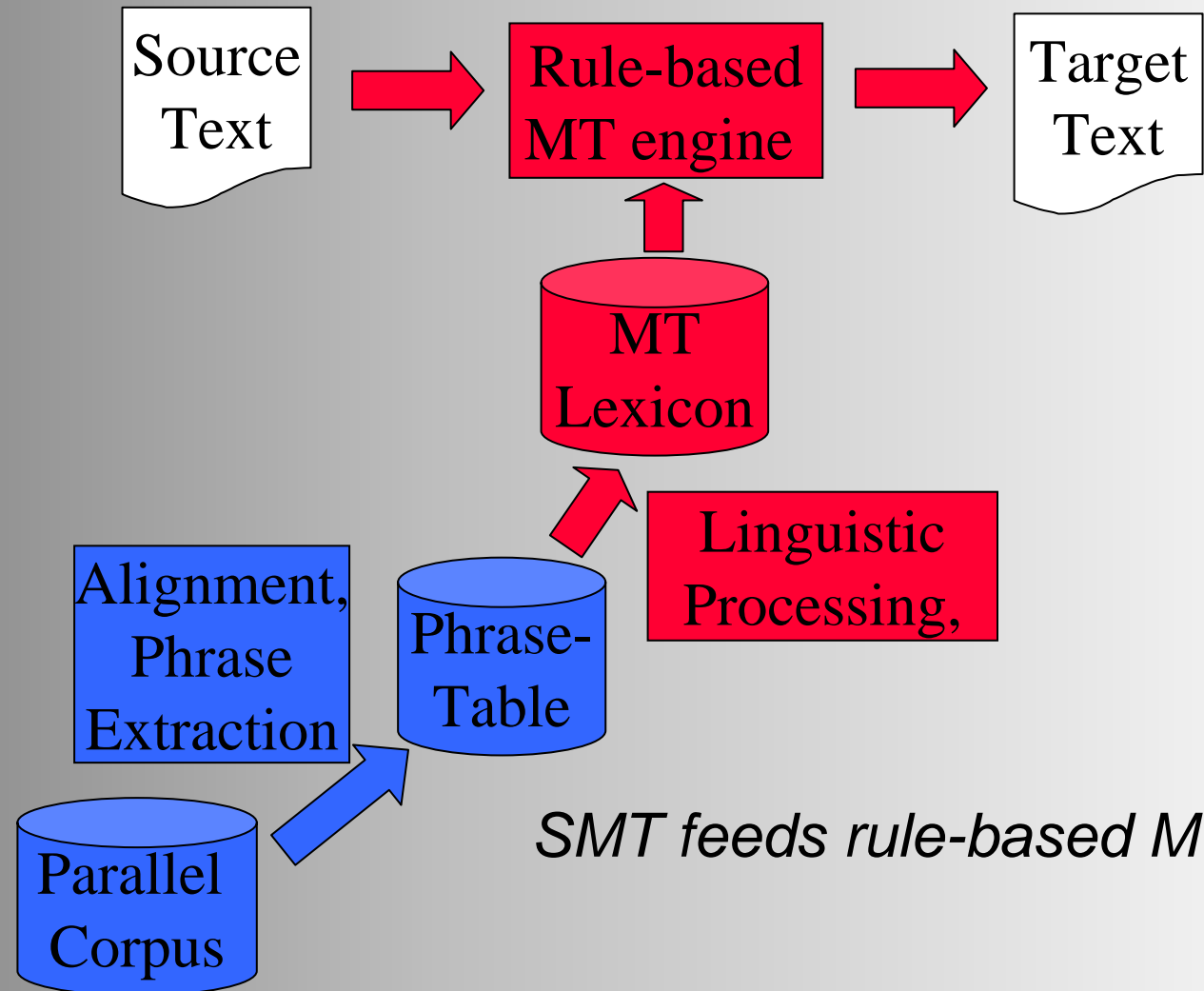
PoS-aware
BLEU-1 score

Systems	Overall (%)	Named Entities (%)
R-I	51.36	34.52
R-II	51.34	64.84
SMT	55.55	20.90
Hybrid	55.53	49.53

Ratio of
untranslated
tokens

Systems	Token #
Ref.	2091 (4.21%)
R-I	3886 (7.02%)
R-II	3508 (6.30%)
SMT	3976 (7.91%)
Hybrid	2425 (5.59%)

Hybrid MT Architecture III



Hybrid MT Architecture III

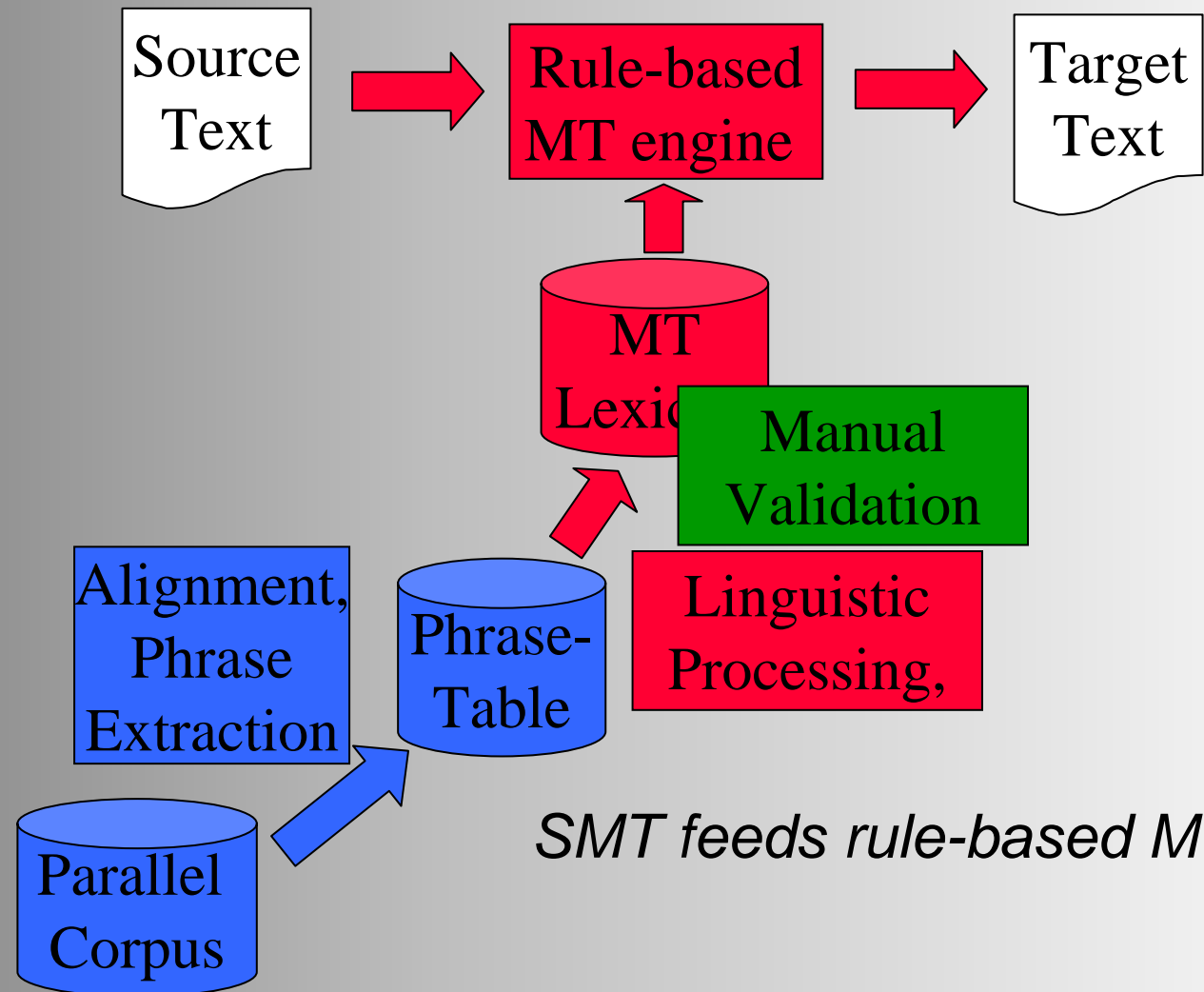


Key issues:

- RBMT has no mechanism to disprefer implausible results → lexicon needs to be correct
- MT lexicon needs information not contained in parallel texts (lemma, gender, inflection, ...)
- Current tools make fully automatic high-quality lexicon extraction rather difficult

Architecture requires manual effort → make it as simple and generic as possible, e.g. by using encoding standard for lexical data like OLIF


Hybrid MT Architecture III



Hybrid MT Architecture III



- OLIF standard has been developed to facilitate exchange of multilingual lexical data.
- Contains encodings for part-of-speech and head, morphosyntactic features, and inflectional behaviour
- Current version 2.1 has focus on English, French, German, Portuguese, and Spanish
- See www.olif.net for details




olif

The OLIF Consortium

Established in March 2000, the OLIF Consortium had its first meeting on April 6, 2000 at SAP headquarters in Walldorf, Germany. The purpose of the consortium, which is headed by SAP, is to generate and support a lexical/terminological interchange standard based on the OLIF prototype (Open Lexicon Interchange Format).

The OLIF prototype was developed within the framework of the OTELO project and distinguished itself from similar exchange formats by providing for both terminological data and the grammatical information required by NLP applications such as MT.

OLIF Consortium members are:



The logos of the OLIF Consortium members are arranged in a grid. The top row includes the European Union flag, Microsoft, Lionbridge, and SAP. The second row features SDL International (with the tagline 'e-enabling global business'), TRADOS, and SYSTRAN (with the tagline 'TRANSLATION SOFTWARE'). The third row shows SAP, acrolinx (with the tagline 'TECHNOLOGY FOR INFORMATION QUALITY'), and the logo of the Institute of Information Technology at the University of Saarland. The bottom row contains IBM, DFKI, the European Patent Office, and linguattec (with the tagline 'language technologies').

OLIF Content
OLIF Home
Documentation
OLIF Tools
Media Center
OLIF Consortium

Downloads
OLIF v.2 DTD
OLIF v.2.1.1
XSD
OLIF v.2 OJPT
Data Archive

Main Sponsor
SAP

Related Links
OLIF Users Group
LISA
AMTA
EAMT
Other Projects

Terminology Extraction from Patents



*Ongoing cooperation between DFKI and
European Patent Office (EPO)*

- Goal: Extract parallel terminologies for EN, DE, ES, FR from translated patent documents
- Motivation for EPO: Build up infrastructure for machine translation of patents, acquire relevant technical vocabulary
- Motivation for DFKI: Develop industrial applications of techniques from MT research, strengthen NLP tools

Terminology Extraction: Motivation



- **Technical documentation** makes up a **large share** of language industry's raw material, vocabulary is **commercially interesting**
- **Manual construction** of broad-coverage or unrestricted multilingual terminologies would be **prohibitively expensive**
- Translated documents exist in large volumes, as well as techniques for sentence/word/phrase alignment, these can be used to partially automate the task
- **IPC** (hierarchical system of about 70K classes) may help to relate extracted terms with **ontologies**
- Test-bed for **scalability** of tools and resources
 - How well do our tools cover technical texts?
 - Can we acquire new lexical information from data?
- First **step towards MT** for technical documents

Terminology Extraction from Patents



History and current status:

- Techniques were prototypically implemented in a feasibility study for WIPO ('03, via acrolinx GmbH)
- Call for Tender by EPO in August '05, bids and results on test data due in September
- From 14 bids, DFKI delivered best results for DE↔EN, ES↔EN and among the best for FR↔EN
- Test phase December '05..July '06: Term extraction from samples, feasibility study on validation
- Production phase (since August '06): Work on 50 million sentence pairs (~ 2E9 running words), manual validation of specific subsets
- Continuation in 2007 may broaden scope to additional languages: PT, IT, RO, NL, SW

Terminology Extraction from Patents



The International Patent Classification (IPC)

- based on the Strasbourg Agreement (1971)
used by >100 national authorities
- indispensable for finding prior art
- hierarchical structure, consisting of
 - eight sections (A..H)
 - 120 classes (A01 ... H05)
 - 628 subclasses (A01B...H05K)
 - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 8th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities

Terminology Extraction from Patents



The International Patent Classification

- based on the Strasbourg Agreement, used by >100 national authorities
- indispensable for finding prior art
- hierarchical structure, consisting of
 - eight sections (A..H)
 - 120 classes (A01 ... H05)
 - 628 subclasses (A01B...H05K)
 - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 8th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities

A: human necessities
B: performing operations; transporting
C: chemistry; metallurgy
D: textiles; paper
E: fixed constructions
F: mechanical engineering; lighting; heating;
G: physics [weapons; blasting
H: electricity

Terminology Extraction from Patents



The International Patent Classification

- based on the Strasbourg Convention, used by >100 nations
- indispensable for finding relevant prior art
- hierarchical structure
 - eight sections (A...H)
 - 120 classes (A01 ... H05)
 - 628 subclasses (A01B...H05K)
 - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 8th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities

A 01 AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING

A 01B SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL

A 01 B 1/00 Hand tools

A 01 B 1/02 spades, shovels

heating;
asting

Terminology Extraction from Patents



The International Patent Classification

- based on the Strasbourg Agreement, used by >100 national authorities
- indispensable for finding relevant prior art
- hierarchical structure
 - eight sections (A...H)
 - 120 classes (A01 ... H05)
 - 628 subclasses (A01B...H05K)
 - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 8th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities

A 01 AGRICULTURE; FORESTRY; ANIMAL

H 05 ELECTRIC TECHNIQUES NOT OTHERWISE PROVIDED FOR

H 05 K PRINTED CIRCUITS; CASINGS OR CONSTRUCTIONAL DETAILS OF ELECTRIC APPARATUS; MANUFACTURE OF ASSEMBLAGES OF ELECTRICAL COMPONENTS

H 05 K 10/00 Arrangements for improving the operating reliability of electronic equipment, e.g. by providing a similar stand-by unit

Terminology Extraction from Patents



Research questions related to the IPC

- Automatic Classification

 - Can IPC classes be identified automatically?

 - (So far classification and search done by ~ 6500 experts)*

- Ontology construction

 - How does the IPC relate to the terminologies used in the various domains? Can we (semi-) automatically construct/extend these terminologies given the documents?

- Word sense disambiguation

 - Can a given IPC class help to identify meaning/translation of a given term?

Terminology Extraction from Patents



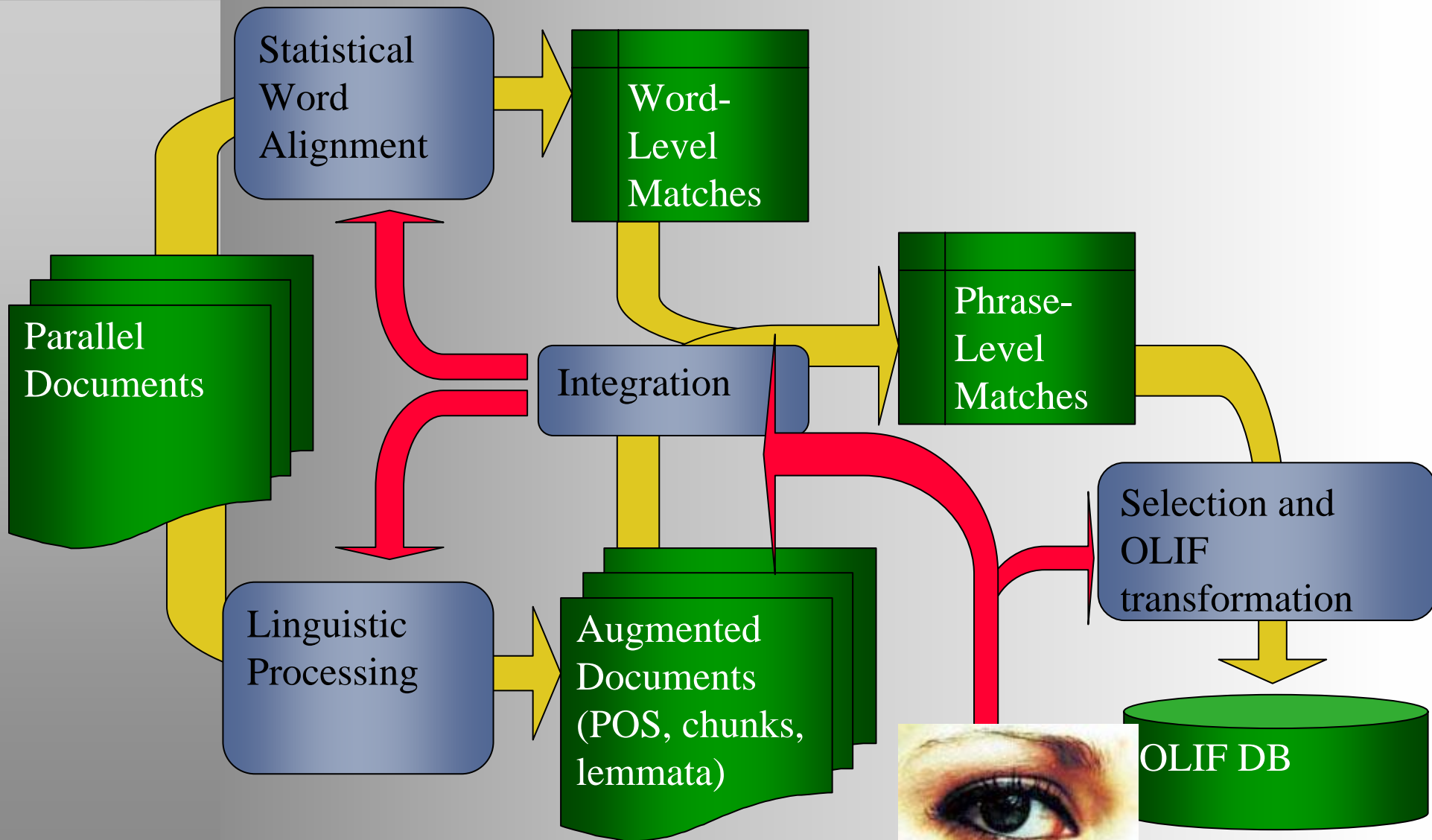
Technical setup:

- Use linguistic tools for corpus annotation
 - POS-tagging, phrase recognition, lemmatization
- Use statistical tools for alignment
 - GIZA++ from Franz Och
 - Own algorithms based on word similarities
- Integrate module outcomes, transform into OLIF entries

Improvement in 2nd phase:

- Feed-back of modifications to basic modules
- Infrastructure for manual validation
- Manual inspection and error analysis is used to improve algorithms as long as the project is ongoing

Terminology Extraction: Architecture



Examples for Patent Terminology



Postbestimmungsortinformationsspeichereinrichtung

= mail destination information memory means

Informationsdurchforstungssteuerungseinrichtung

= information browsing control means

Hypervideonachrichtversendungsverarbeitungseinrichtung

= hypervideo message posting processing means

Gasphasenverunreinigungsabsorptionsflüssigkeit

= gas phase contaminant absorbing liquid

Manual Validation of Terminology



- Original Plan:
 - Validation by (30..40) domain experts in national patent offices, but:
 - Linguistic validation not suitable for patent examiners
- New setup: Validation work is shared between
 - DFKI for linguistic validation and
 - patent offices for domain knowledge
- Validation workflow handled in a Web-based infrastructure for terminology maintenance
 - Prototype available since Fall '06
 - Successfully used for first deliveries

Manual Validation of Terminology



Mozilla Firefox

http://termvalidation2.dfki.de/pairs/monodetailsDE/val:%3Cnew/0/ "work in progress"

DE Search: val:<new (26588 entries)

Search: val:<new DE Go up prev next Configure Presentation

CanForm	PoS	Head	Number	Gender	genSg	nomPl	datPl	Paradigm	Delete?	Accept	Validator
A Laboratory	noun	Laboratory	sg	m				\$	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Dorothea
A Laboratory Manual	noun	Manual	sg	n	Manual(e)s	Manuals	Manuals	Tank	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
A-Block	noun	Block	sg	m	Block(e)s	Blöcke	Blöcken	Wunsch	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
A-Eingang	noun	Eingang	sg	m	Eingang(e)s	Eingänge	Eingängen	Wunsch	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
A-Einheit	noun	Einheit	sg	f	Einheit	Einheiten	Einheiten	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
A-Kette	noun	Kette	sg	f	Kette	Ketten	Ketten	Feder	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
A-Phase	noun	Phase	sg	f	Phase	Phasen	Phasen	Feder	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
A-Register	noun	Register	sg	n	Registers	Register	Registern	Splitter	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
A-Seite	noun	Seite	sg	f	Seite	Seiten	Seiten	Feder	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
ABS-Steuerung	noun	Steuerung	sg	f	Steuerung	Steuerungen	Steuerungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AC-Generator	noun	Generator	sg	m	Generators	Generatoren	Generatoren	Traktor	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
ACE-Inhibitor	noun	Inhibitor	sg	m	Inhibitors	Inhibitoren	Inhibitoren	Traktor	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AFC-Schaltung	noun	Schaltung	sg	f	Schaltung	Schaltungen	Schaltungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AGC-Schaltung	noun	Schaltung	sg	f	Schaltung	Schaltungen	Schaltungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AGC-Spannung	noun	Spannung	sg	f	Spannung	Spannungen	Spannungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AGC-Verstärker	noun	Verstärker	sg	m	Verstärkers	Verstärker	Verstärkern	Splitter	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AGR-Ventil	noun	Ventil	sg	n	Ventils	Ventile	Ventilen	Ziel	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AIDS Virus	noun	Virus	sg	n	Virus	Viren	Viren	Zyklus	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AND Gate	noun	Gate	sg	n	Gates	Gates	Gates	Code	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AND Schaltung	noun	Schaltung	sg	f	Schaltung	Schaltungen	Schaltungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AND-Gatter	noun	Gatter	sg	n	Gatters	Gatter	Gattern	Splitter	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AND-Glied	noun	Glied	sg	n	Glied(e)s	Glieder	Gliedern	Leib	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
ANDF-Installationseinri	noun	Einrichtung	sg	f	Einrichtung	Einrichtungen	Einrichtungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
ANDF-Herstellungseinri	noun	Einrichtung	sg	f	Einrichtung	Einrichtungen	Einrichtungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
AO-Katalysator	noun	Katalysator	sg	m	Katalysators	Katalysatoren	Katalysatoren	Traktor	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
APC-Schaltung	noun	Schaltung	sg	f	Schaltung	Schaltungen	Schaltungen	Frau	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
ART	noun	ART	sg	f				\$	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Dorothea
ASTMes-Verfahren	noun	Verfahren	sg	n	Verfahrens	Verfahren	Verfahren	Wagen	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
ATM-Netz	noun	Netz	sg	n	Netzes	Netze	Netzen	Tisch	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea
ATM-Netzwerk	noun	Netzwerk	sg	n	Netzwerk(e)s	Netzwerke	Netzwerken	Termin	<input type="checkbox"/>	<input type="checkbox"/>	Dorothea

submit+next Anfrage abschicken Zurücksetzen up prev next

You are logged in as eisele [logout](#) You have already edited 1690 entries!

Hybrid MT Architecture III



Next steps:

- Use existing infrastructure to feed various rule-based MT engines (OpenLogos, Lucy)
- Measure impact on results
- Decide on domain for which extended lexicons would be most useful

Conclusion



- We have presented two complementary architectures to combine rule-based and statistical MT engines
- Implementation is fairly advanced but fine-tuning still needs to be done
- These setups can themselves be combined into a MEMT system
- Truly deep integration using rule-based and statistical **knowledge sources** in well-balanced way will need more work

Thank You for Your Attention

