



From Statistical Term Extraction to Hybrid Machine Translation

Petra Wolf, Ulrike Bernardi
Lucy Software and Services, Munich

Christian Federmann, Sabine Hunsicker
DFKI, Saarbrücken

Contents

- Motivation
- Intelligent Terminology Extraction for Hybrid MT
- Evaluation
- Conclusion

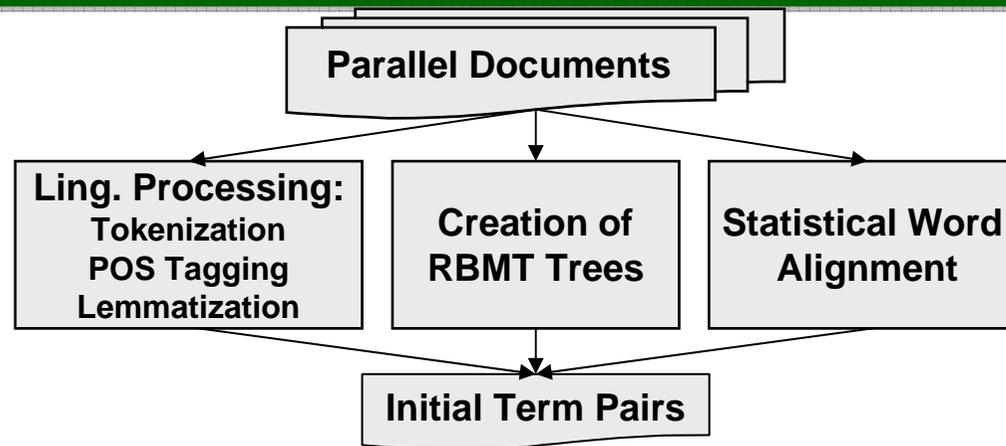
- SMT: superior selection of target entities
- Phrase table access at RBMT transfer: strong & weak
- Deeper intertwined hybrid extension

⇒ **L i S T E X:**

Hybrid Transfer by

Linguistically augmented Statistical Terminology EXtraction

Intelligent Terminology Extraction

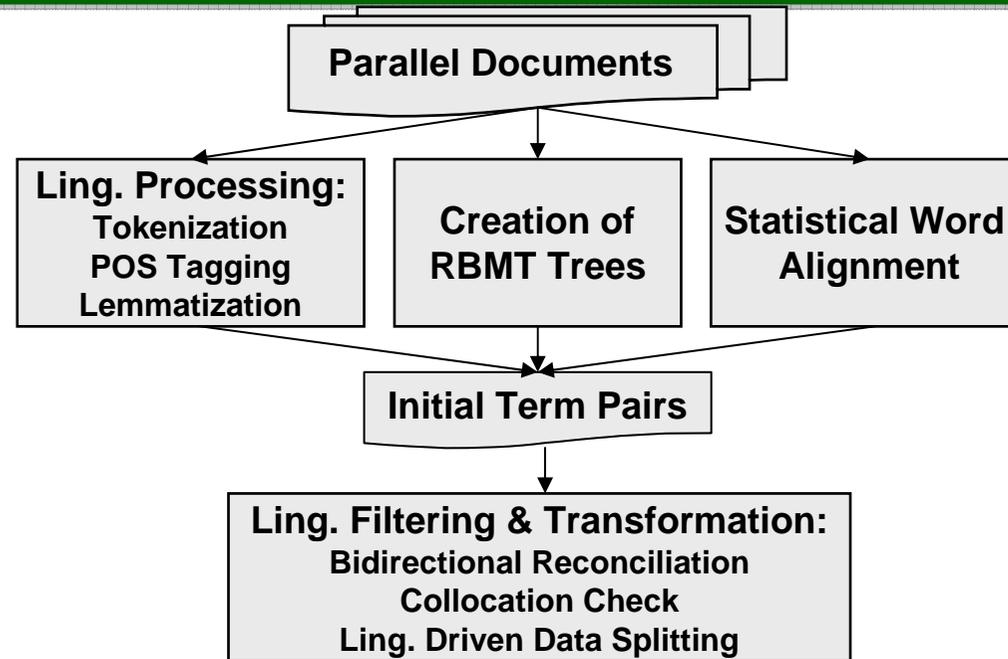


Initial Term Pairs - Examples

Abfallverbrennungsanlage; NST; waste incineration plant; NST;3;
Die Technologie zur Rauchgasreinigung in Abfallverbrennungsanlagen
existiert bereits .; The technology for cleaning up fumes from waste
incineration plants already exists .;NST AST; NST NST

Beitrittsverhandlung;NST;accession negotiation;NST;208;
Malta tritt in die Beitrittsverhandlungen ein .;
Malta is entering the accession negotiations .;NST;NST NST

Intelligent Terminology Extraction



Reduction of false entries:

- Extract frequencies (for single and multiwords)
- Reject single terms with lower/equal frequency than corresponding multiwords

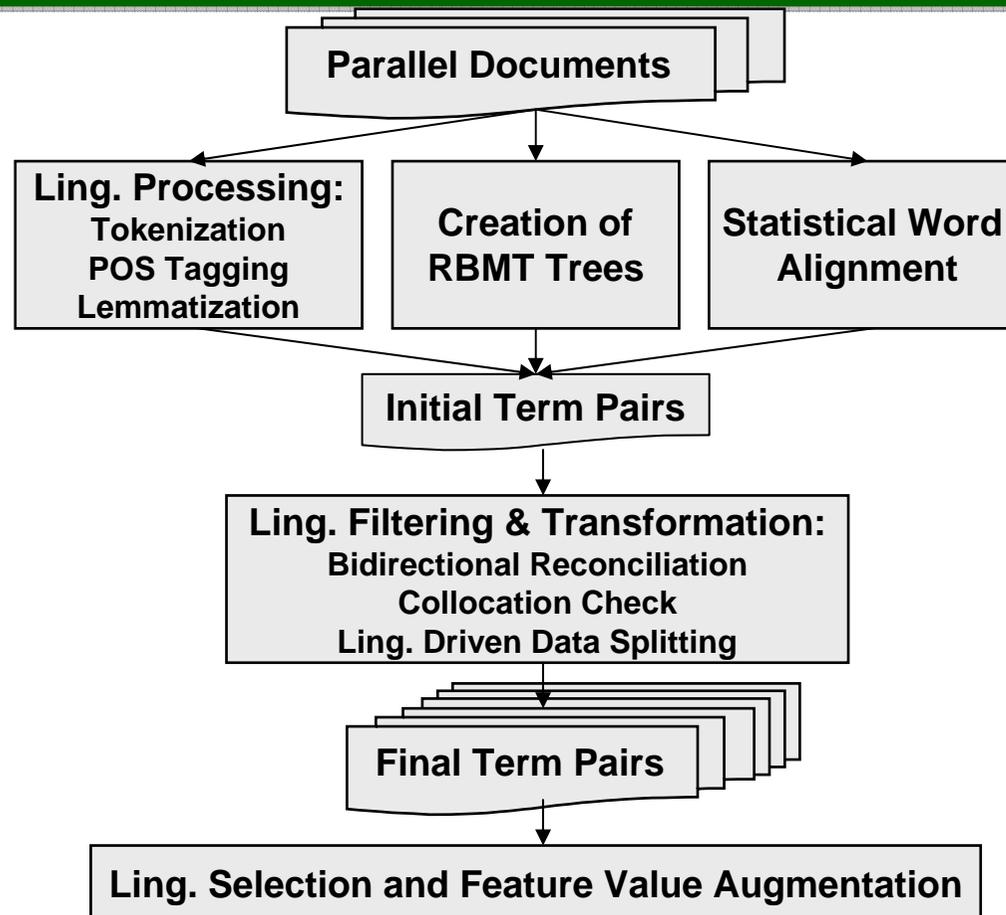
Tour de France → Tour de France - Tour de France

→ ~~France - France~~

Higher precision by data splitting in case of:

- Lemmatizer does not find correct lemma and just guesses it
- Special characters, e.g. punctuation or integers
- Category changes
- Quality Splitting:
 1. Single words on source as well as on target side
 2. Single words as source, multiword expressions as targets
 3. Multiword expressions as source, single words as targets
 4. Multiword expressions on source as well as on target side
- POS distinction for nouns, verbs, adjectives, adverbs

Intelligent Terminology Extraction



- Quality Precision & Selection:
 - Resolve lemmatizer and derivational errors
 - Rectify wrong or missing endings
 - Delete senseless terms, such as *of of*
 - Delete wrongly categorized entries
 - Evaluate nouns in plural form
 - Delete wrong translation equivalents with certain syntax
- Feature Value Augmentation
 - Single Word Defaulting
 - Multiword Defaulting
 - Defaulting Monolingual Entries from Transfer
 - Deriving

Example for Feature Value Augmentation:

Beitrittsverhandlung – accession negotiation

- Defaulted bilingual entry:

*("Beitrittsverhandlung" NST "accession negotiation" NST PRF 10000
TAG (POL) !DATE 1296734170 !OWNER "sys" !AUTHOR
"TermExtract")*

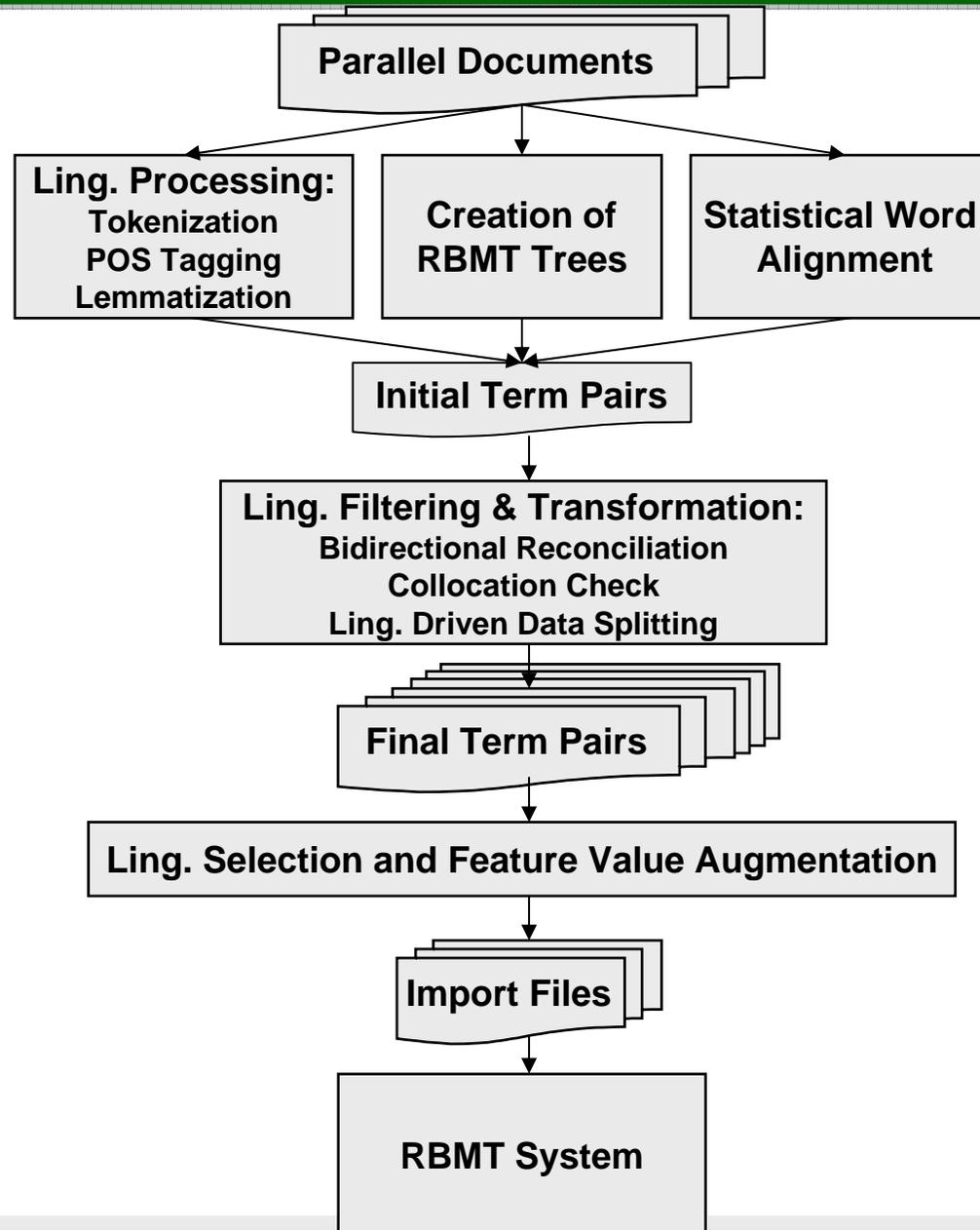
- Defaulted German monolingual entry:

*("Beitrittsverhandlung" NST ALO "Beitrittsverhandlung" ARGS ((N1
(PREP "mit" "über") (CA A)) (N0 (PREP0 "über") (FCP TH) (INT T))) CL
(P-EN S-0) GD (F) KN MS-CNT LINK S SX (N) TYN (PRO) !AUTHOR
"TermExtract" !OWNER "sys" !DATE 1296734190)*

- Defaulted English monolingual entry:

*("accession negotiation" NST ALO "accession negotiation" KN CNT
TYN (ABS PRO) MW-TYPE STRING-NST MW-BODY ((STRING
"accession") (HEAD)) !AUTHOR "TermExtract" !OWNER "sys" !DATE
1296734198 MW-HEAD-CAN "negotiation" MWHEAD-CAT NST)*

Intelligent Terminology Extraction



Translation Direction	German-English	Spanish-English
Lines in Europarl Corpus	1,259,571	1,253,026
Monolingual Lists	441,425 / 508,592	406,296 / 294,069
Initial Term Pairs	45,857	35,088
Final Term Pairs	30,803	24,519
Imported Terminology	27,054	18,546

Improvements:

■ Superior selection of target entities:

Source:	<i>prostitución infantil</i>
LT-BASE:	<i>infantile prostitution</i>
LiSTEX:	<i>child prostitution</i>

■ Better recognition of whole sentence structure:

Source: ... *but I must stress that in some Member States the collection of this kind of data infringes constitutional provisions on the protection of privacy.*

LT-BASE: ... *aber ich betonen muss, dass die Sammlung dieser Art von Daten in einigen Mitgliedstaaten Verfassungsbereitstellungen auf dem Schutz von Privatleben verletzt.*

LT-LiSTEX: ... *aber ich muss betonen, dass die Sammlung dieser Art von Daten in einigen Mitgliedstaaten verfassungsrechtliche Bestimmungen auf dem Schutz der Privatsphäre verletzt.*

Multiword Parts Added and/or Lost:

- Alignment errors:

 - Source: *Zeitraum*

 - LT-BASE: *period*

 - LT-LiSTEX: [*five-year period|period*]

- Free or elliptic translations in corpus:

 - Source: *Rechnungshof*

 - LT-BASE: *court of auditors*

 - LT-LiSTEX: *Court*

*Der Rechnungshof arbeitet jedoch mit Normen. →
Nonetheless, the Court is working with standards.*

Multiword Expressions and Collocations:

- Large syntactic variability:

Source: *des Europäischen Rates in Nizza und in Biarritz*

LT-BASE: *of the European [Council|Advice] in Nice and in Biarritz*

LT-LISTEX: *of the Nice Council and in Biarritz*

→ Extend intelligent representation and grammar processing for variable collocational syntax.

Wrong Capitalization of English Terms:

- Capitalized vs. non-capitalized translation alternatives:

Source: ... *die Verfassung*

LT-BASE: ... *the [constitution|state]*

LT-LISTEX: ... [*Constitution|constitution*]

→ Only most frequent spelling variant will remain.

Wrong Translation Equivalents:

- Wrong translations in the original corpus:

November → October

mañana → afternoon

- Idiomatic usage in the original corpus:

Gleichwohl dürfte es schwierig sein, das Rad zurückzudrehen. →

However, the impression is that it will be difficult to turn the clock back.

→ *Rad → clock:*

Der Sieg von Paolo Bettini bei der Rad-Weltmeisterschaft → The victory of Paolo Bettini during the clock world championship

- Only very few examples

Missing Subcategorization Frames:

- *Bemerkung* zu jdn./etw. → *remark* on sb./sth.
- Extracted terminology does not include frame information.

→ Differentiate merge of information from existing lexicon entries and new terms.

Translation Quality Evaluation

	English ↔ Spanish		English ↔ German	
	English → Spanish	Spanish → English	English → German	German → English
Translated TUs	2,000	2,000	2,000	2,000
Different Translations	95.05%	96.05%	95.45%	96.20%
Evaluated Differences	287	398	970	1,261
Better	47.74%	38.44%	50.82%	57.49%
Equal	31.71%	46.73%	41.86%	30.29%
Worse	20.56%	14.82%	7.32%	12.21%

Result:

- Great potential in multi-level hybrid approach of LiSTEX
- Viable hybrid system application
- Promising quality gains

Outlook and next steps:

- LiSTEX in other domains / corpora
- Lemmatizer and generation improvements
- Intelligent handling of syntactically free collocations
- Refine translation alternatives' selection



Thank you for your attention.

Translation Quality Evaluation

Translation Direction	Baseline		LiSTEX	
	NIST	BLEU	NIST	BLEU
German → English	5.5582	0.1632	5.2430	0.1491
English → German	4.3616	0.1078	4.2718	0.1060
Spanish → English	5.8776	0.1953	5.6414	0.1830
English → Spanish	6.1375	0.2085	5.9086	0.1978