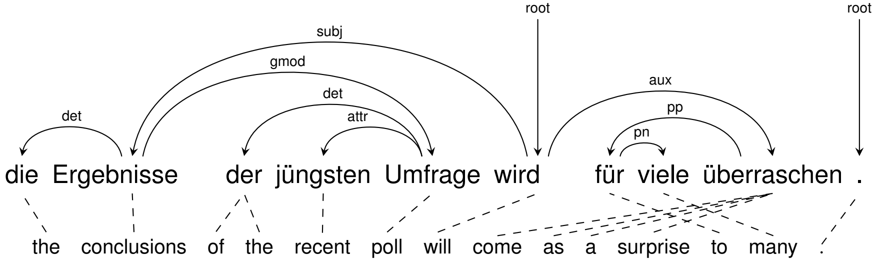
Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation

Rico Sennrich

Institute for Language, Cognition and Computation
University of Edinburgh

September 19 2015

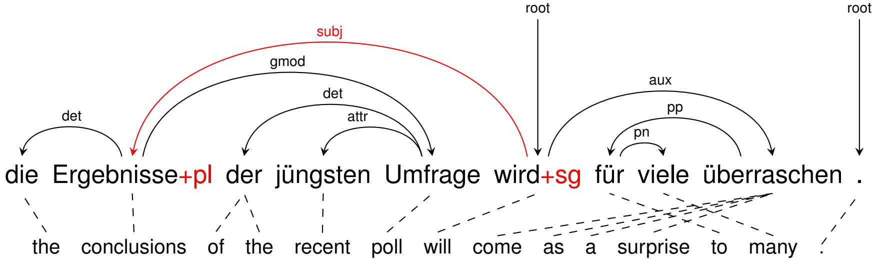
Problem: ungrammatical translation output



what's wrong?

- subject-verb agreement: *die Ergebnisse* (pl) – *wird* (sg)
- subcategorisation: *überraschen* is transitive

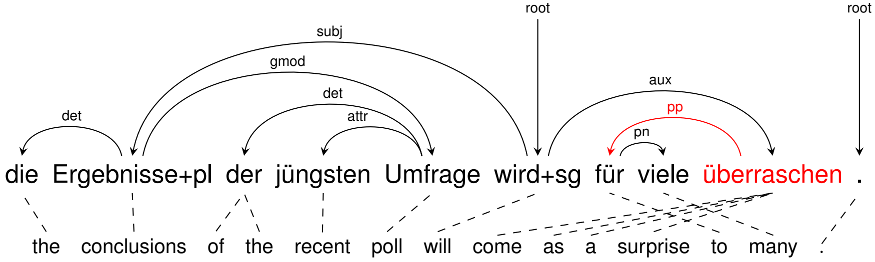
Problem: ungrammatical translation output



what's wrong?

- subject-verb agreement: *die Ergebnisse* (pl) – *wird* (sg)
- subcategorisation: *überraschen* is transitive

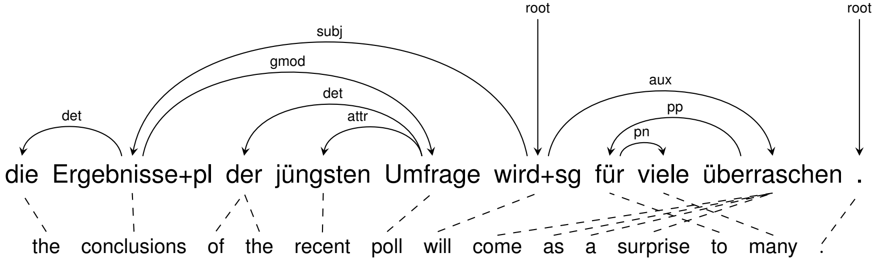
Problem: ungrammatical translation output



what's wrong?

- subject-verb agreement: *die Ergebnisse* (pl) – *wird* (sg)
- subcategorisation: *überraschen* is transitive

Problem: ungrammatical translation output



what's wrong?

- subject-verb agreement: *die Ergebnisse* (pl) – *wird* (sg)
- subcategorisation: *überraschen* is transitive

syntactic n-grams

- n-gram language models are sensitive to string distance
- dependency chains (rebranded **syntactic n-grams** [Sidorov et al., 2013]) are more robust

Contribution

previous work

- large body of research on syntactic language models for SMT
  [Charniak et al., 2003, Och et al., 2004, Quirk et al., 2004, Post and Gildea, 2008, Cherry and Quirk, 2008, Shen et al., 2010]
- promising results with dependency language models

our contribution

- novel **relational** dependency language model
- optimization of global SMT parameters on syntactic MT metric
  $\rightarrow$ better appreciation of syntactic language models

Towards a relational dependency language model

previous work [Quirk et al., 2004, Shen et al., 2010]

- unlabelled
- varying degrees of word order modeling:
  - none [Quirk et al., 2004]
  - heavy reliance on position [Shen et al., 2010]

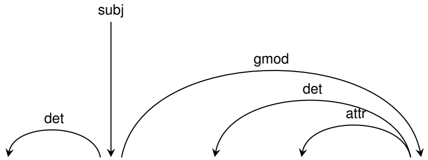
our model

- relational: dependency labels as atomic elements
  - use dependency labels as context
    verb must agree with subject, but not with object
  - also predict dependency labels
    side-effect: models subcategorisation
- sibling order is considered, but not relied on

Notation

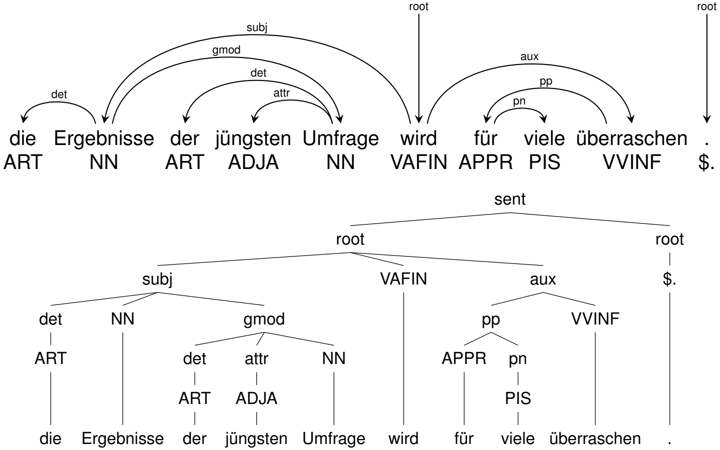
- S: sequence of words
- D: sequence of dependency labels
- T: sequence of head positions (tree topology)

common approximation: $P(S) \approx P(S|T)$



S	die	Ergebnisse	der	jüngsten	Umfrage
D	det	subj	det	attr	gmod
T	2	0	5	5	2

Side note: conversion to constituency format

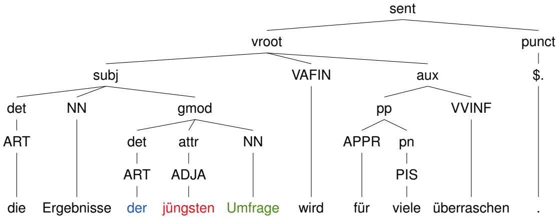


Dependency Language Model (DLM)

$$P(S) = P(w_1, w_2, ..., w_n) \approx \prod_{i=1}^{n} P(w_i | h_s(i), h_a(i)) \quad (1)$$

Markov assumption: use window of (closest) q siblings and r ancestors:

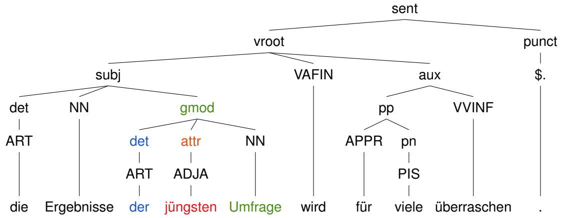
$$P(S) \approx \prod_{i=1}^{n} P(\color{red}{w_i}|\color{blue}{h_s(i)_1^q}, \color{green}{h_a(i)_1^r}) \quad (2)$$



Relational Dependency Language Model (RDLM)

relational model predicts dependency labels, and is conditioned on ancestor/sibling labels:

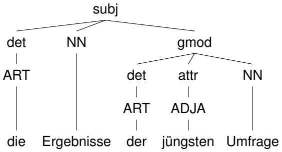
$$
\begin{aligned}
P(S,D) &= P(D) \times P(S|D) \\
&\approx \prod_{i=1}^{n} P_l(i) \times P_w(i) \\
P_l(i) &= P(l_i \;| h_s(i)_1^q, l_s(i)_1^q, h_a(i)_1^r, l_a(i)_1^r) \\
P_w(i) &= P(w_i \;| h_s(i)_1^q, l_s(i)_1^q, h_a(i)_1^r, l_a(i)_1^r, l_i)
\end{aligned}
\tag{3}
$$



Predicting Tree Topology

final model generates all (m) nodes, including preterminals (<PT>) and virtual STOP nodes (<S>).

$$P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases} \quad (4)$$

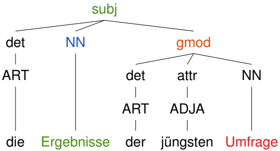


N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
D	subj	det	<PT>	<S>	<PT>	gmod	det	<PT>	<S>	attr	<PT>	<S>	<PT>	<S>	<S>
S	Ergebnisse	die	ϵ	ϵ	ϵ	Umfrage	der	ϵ	ϵ	jüngsten	ϵ	ϵ	ϵ	ϵ	ϵ
T	0	1	2	2	1	1	6	7	7	6	10	10	6	6	1

Predicting Tree Topology

final model generates all (m) nodes, including preterminals (<PT>) and virtual STOP nodes (<S>).

$$P(S,D,T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases} \quad (4)$$

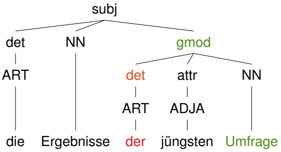


N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
D	subj	det	<PT>	<S>	<PT>	gmod	det	<PT>	<S>	attr	<PT>	<S>	<PT>	<S>	<S>
S	Ergebnisse	die	ϵ	ϵ	ϵ	Umfrage	der	ϵ	ϵ	jüngsten	ϵ	ϵ	ϵ	ϵ	ϵ
T	0	1	2	2	1	1	6	7	7	6	10	10	6	6	1

Predicting Tree Topology

final model generates all (m) nodes, including preterminals (<PT>) and virtual STOP nodes (<S>).

$$P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases} \tag{4}$$

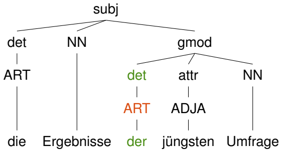


N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
D	subj	det	<PT>	<S>	<PT>	gmod	det	<PT>	<S>	attr	<PT>	<S>	<PT>	<S>	<S>
S	Ergebnisse	die	ϵ	ϵ	ϵ	Umfrage	der	ϵ	ϵ	jüngsten	ϵ	ϵ	ϵ	ϵ	ϵ
T	0	1	2	2	1	1	6	7	7	6	10	10	6	6	1

Predicting Tree Topology

final model generates all (m) nodes, including preterminals (<PT>) and virtual STOP nodes (<S>).

$$P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases} \tag{4}$$

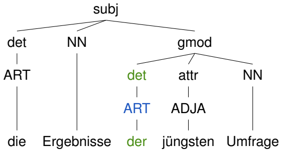


N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
D	subj	det	<PT>	<S>	<PT>	gmod	det	<PT>	<S>	attr	<PT>	<S>	<PT>	<S>	<S>
S	Ergebnisse	die	ϵ	ϵ	ϵ	Umfrage	der	ϵ	ϵ	jüngsten	ϵ	ϵ	ϵ	ϵ	ϵ
T	0	1	2	2	1	1	6	7	7	6	10	10	6	6	1

Predicting Tree Topology

final model generates all (m) nodes, including preterminals (<PT>) and virtual STOP nodes (<S>).

$$P(S,D,T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases} \quad (4)$$

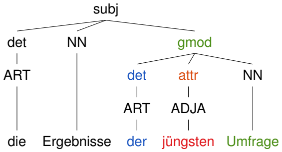


N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
D	subj	det	<PT>	<S>	<PT>	gmod	det	<PT>	<S>	attr	<PT>	<S>	<PT>	<S>	<S>
S	Ergebnisse	die	ϵ	ϵ	ϵ	Umfrage	der	ϵ	ϵ	jüngsten	ϵ	ϵ	ϵ	ϵ	ϵ
T	0	1	2	2	1	1	6	7	7	6	10	10	6	6	1

Predicting Tree Topology

final model generates all (m) nodes, including preterminals (<PT>) and virtual STOP nodes (<S>).

$$P(S,D,T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases} \qquad (4)$$



N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
D	subj	det	<PT>	<S>	<PT>	gmod	det	<PT>	<S>	attr	<PT>	<S>	<PT>	<S>	<S>
S	Ergebnisse	die	ϵ	ϵ	ϵ	Umfrage	der	ϵ	ϵ	jüngsten	ϵ	ϵ	ϵ	ϵ	ϵ
T	0	1	2	2	1	1	6	7	7	6	10	10	6	6	1

Training

Neural Network Training

- feed-forward network architecture similar to [Vaswani et al., 2013]
- separate networks for P_l and P_w
- one hidden layer
- big vocabulary: 500 000

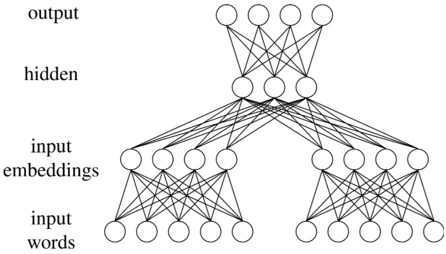


Figure : Neural network architecture [Vaswani et al., 2013]

Decoding

Decoding with (R)DLM

- string-to-tree SMT decoder
  - decoder builds dependency trees
  - we score each hypothesis with (R)DLM
- decoding is bottom-up, but (R)DLM is top-down
  - dummy tokens for unavailable context
  - embedding of dummy token is weighted average of all words/labels
  - nodes are rescored as more context becomes available

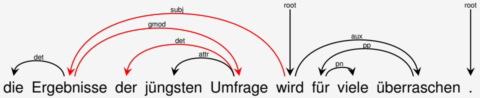
A syntactic SMT metric for optimization and evaluation

Desideratum

- metric that rewards grammaticality beyond n-grams

Head-word chain metric (HWCM) [Liu and Gildea, 2005]

- precision-oriented reference-based metric (like BLEU)
- precision is estimated for dependency chains instead of n-grams



example chain: wird - Ergebnisse - Umfrage - der

Our contribution

- we use HWCM (f-score) for optimization of SMT parameters.
  $\rightarrow$ first use of (non-shallow) syntactic metric for tuning

Evaluation

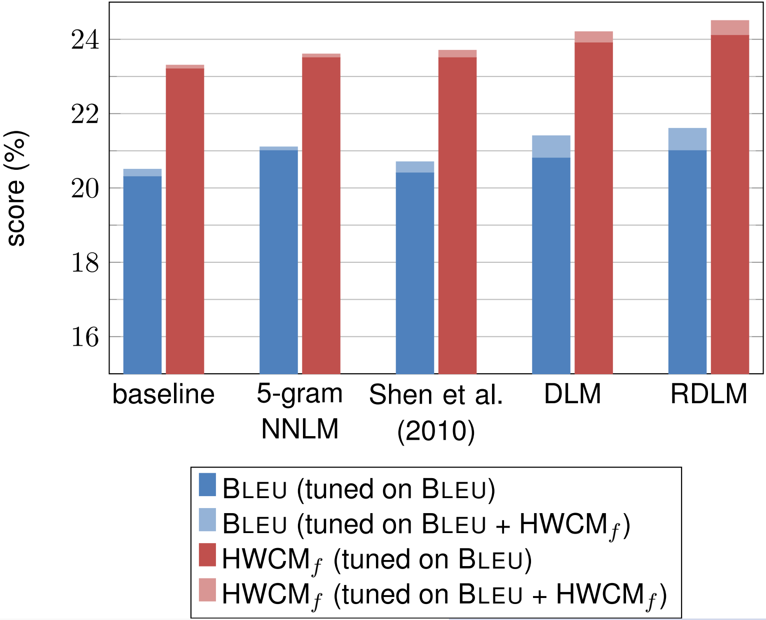
Metrics

- automatic SMT metrics
- agreement errors

Data and methods

- English-German (and -Russian) data from WMT 2014
- 4.5 million sentence pairs parallel data; 120 million sentences monolingual data
- automatically parsed with ParZu [Sennrich et al., 2013]
- string-to-tree baseline as in [Williams et al., 2014]
- 3 runs of k-best batch MIRA optimization
- Moses toolkit

Evaluation: English→German (newstest2014)



Evaluation: automatic SMT metrics (newstest2014)

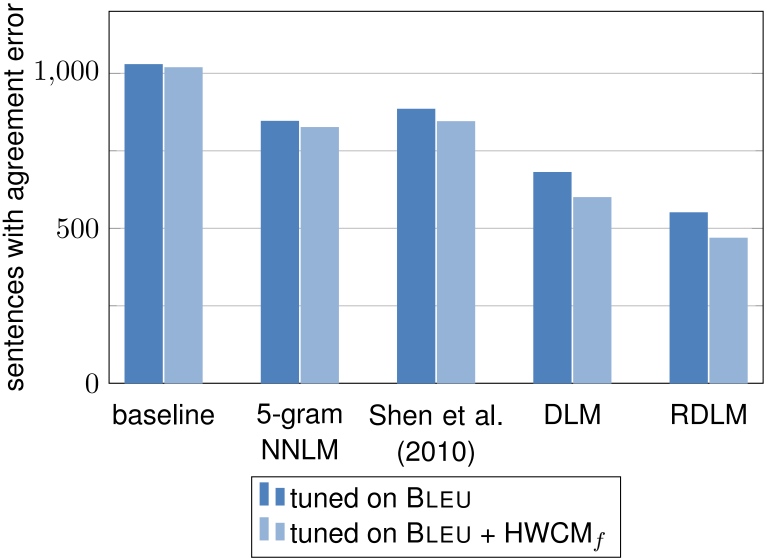
English→German

system	BLEU	HWCM_f
baseline	20.3	23.2
+RDLM	21.0	24.1
+HWCM tuning	21.6	24.5

English→Russian

system	BLEU	HWCM_f
baseline	25.9	23.9
+RDLM	26.6	26.5
+HWCM tuning	26.8	27.3

Evaluation: morphological agreement errors



Conclusions

relational dependency language model (RDLM)

- substantially improves fluency
  (BLEU/HWCM$_f$; agreement errors; ranked 1–2 (out of 16) @ WMT 15)
- relational variant outperforms unlabelled model and related work

HWCM tuning

- dependency-based metric suitable for tuning
  (see also: RED @ WMT15 tuning task)
- synergy effects between metric and model

follow-up work

A Joint Dependency Model of Morphological and Syntactic Structure for SMT come see my talk! (Mo, 13:45, room 1)

Thank you!

code

- RDLM/HWCM are integrated in Moses: `http://statmt.org/moses/`
- configs: `https://github.com/rsennrich/wmt2014-scripts`

Bibliography I

Charniak, E., Knight, K., and Yamada, K. (2003).
Syntax-based language models for statistical machine translation.
In MT Summit IX, New Orleans, USA.

Cherry, C. and Quirk, C. (2008).
Discriminative, Syntactic Language Modeling through Latent SVMs.
In Proceedings of AMTA 2008.

Fraser, A., Weller, M., Cahill, A., and Cap, F. (2012).
Modeling Inflection and Word-Formation in SMT.
In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 664–674, Avignon, France. Association for Computational Linguistics.

Liu, D. and Gildea, D. (2005).
Syntactic Features for Evaluation of Machine Translation.
In
Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 25–32, Ann Arbor, Michigan.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004).
A Smorgasbord of Features for Statistical Machine Translation.
In
Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association
pages 161–168, Boston, Massachusetts, USA. Association for Computational Linguistics.

Post, M. and Gildea, D. (2008).
Parsers as language models for statistical machine translation.
In Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas.

Bibliography II

Quirk, C., Menezes, A., and Cherry, C. (2004).
Dependency Tree Translation: Syntactically Informed Phrasal SMT.
Technical Report MSR-TR-2004-113, Microsoft Research.

Rosa, R., Mareček, D., and Dušek, O. (2012).
DEPFIX: A System for Automatic Correction of Czech MT Outputs.
In Proceedings of the Seventh Workshop on Statistical Machine Translation, WMT '12, pages 362–368, Montreal, Canada. Association for Computational Linguistics.

Sennrich, R., Volk, M., and Schneider, G. (2013).
Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis.
In Proceedings of the International Conference Recent Advances in Natural Language Processing 2013, pages 601–609, Hissar, Bulgaria.

Shen, L., Xu, J., and Weischedel, R. (2010).
String-to-dependency Statistical Machine Translation.
Comput. Linguist., 36(4):649–671.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernández, L. (2013).
Syntactic Dependency-based N-grams As Classification Features.
In Proceedings of the 11th Mexican International Conference on Advances in Computational Intelligence - Volume Part II, MICAI'12, pages 1–11, Berlin, Heidelberg. Springer-Verlag.

Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013).
Decoding with Large-Scale Neural Language Models Improves Translation.
In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, pages 1387–1392, Seattle, Washington, USA.

Bibliography III

Williams, P., Sennrich, R., Nadejde, M., Huck, M., Hasler, E., and Koehn, P. (2014).
Edinburgh's Syntax-Based Systems at WMT 2014.
In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 207–214, Baltimore, Maryland, USA. Association for Computational Linguistics.

Evaluation: English→Russian

MIRA objective	system	dev			newstest2013			newstest2014		
		BLEU	HWCM$_f$	TER	BLEU	HWCM$_f$	TER	BLEU	HWCM$_f$	TER
BLEU	baseline	22.5	21.6	56.7	17.1	18.8	64.7	25.9	23.9	54.5
	DLM	**23.3***	23.5	**56.0**	**17.5**	20.2	64.0	26.4	26.1	**53.8**
	RDLM	**23.1**	**23.7**	**56.0**	**17.6**	**20.4**	**63.8**	**26.6**	**26.5**	**53.7**
BLEU+ HWCM$_f$	baseline	22.5	22.9*	56.1*	17.2	19.7*	63.9*	25.8	25.1*	54.1*
	DLM	**23.0**	24.1*	**55.6***	**17.6**	**20.8***	**63.2***	26.4	26.9*	53.3*
	RDLM	**23.1**	**24.4***	**55.4***	**17.6**	**20.9***	**63.1***	**26.8***	**27.3***	**53.0***

Table : Translation quality of English→Russian string-to-tree SMT system.

Evaluation: automatic SMT metrics

MIRA objective	system	dev				newstest2013				newstest2014			
		BLEU	HWCM$_f$	METEOR	TER	BLEU	HWCM$_f$	METEOR	TER	BLEU	HWCM$_f$	METEOR	TER
BLEU	baseline	34.4	32.6	52.5	47.4	19.8	22.8	39.7*	62.4	20.3	23.2	42.0*	62.7
	5-gram NNLM	**35.3**	33.1	**53.2***	**46.4**	**20.4**	23.2	40.2	**61.7**	**21.0**	23.5	42.5*	**62.2**
	[Shen et al., 2010]	34.4*	33.2	52.7*	46.9	20.0	23.2	40.0*	62.3	20.4	23.5	42.3*	62.9
	DLM	34.9*	**33.8**	**53.1***	46.8	20.3	23.6	40.1*	**61.7**	20.8	23.9	42.3*	**62.2**
	RDLM	35.0	**33.9**	**53.1***	46.7	**20.5**	**23.8**	**40.4***	**61.7**	**21.0**	**24.1**	**42.7***	**62.2**
	5-gram + RDLM	35.5	34.0	53.4*	46.3	20.7	23.7	40.6*	61.5	21.4	24.1	42.9*	61.7
BLEU + HWCM$_f$	baseline	34.4	33.0*	52.4	46.9*	20.0*	23.0*	39.6	61.9*	20.5*	23.3*	41.8	62.2*
	5-gram NNLM	**35.2**	33.5*	**53.0**	**46.0***	20.6*	23.4*	40.1	60.9*	21.1*	23.6	42.3	61.5*
	[Shen et al., 2010]	34.2	33.8*	52.4	46.4*	20.2*	23.5*	39.8	61.8*	20.7*	23.7*	42.1	62.2*
	DLM	34.8	34.3*	52.7	**45.9***	20.4	23.8*	39.8	**60.7***	21.4*	24.2*	42.0	**60.9***
	RDLM	34.9	**34.5***	**53.0**	**45.8***	**20.9***	**24.2***	**40.3**	**60.7***	**21.6***	**24.5***	**42.5**	**60.8***
	5-gram + RDLM	35.4	34.6*	53.2	45.4*	21.0*	24.1*	40.4	60.5*	21.8*	24.4*	42.7	60.6*

Table : Translation quality of English→German string-to-tree SMT system.

Meta-Evaluation

METEOR	-0.54
BLEU	-0.77
TER	0.69
HWCM$_f$	**-0.92**

System-level rank correlation (Kendall's τ) between automatic metrics and number of agreement errors.

Evaluation: examples

source	also **the user** manages his identity and **can** therefore be anonymous.
baseline	auch **der Benutzer** verwaltet seine Identität und **können** daher anonym sein.
RDLM	auch **der Benutzer** verwaltet seine Identität und **kann** daher anonym sein.
ref	darüber hinaus verwaltet **der Inhaber** seine Identität und **kann** somit anonym bleiben.

subject-verb agreement

baseline has singular subject, but plural verb

Evaluation: example

source	how do you **apply this definition** to their daily life and social networks?
baseline	wie kann man **diese Definition** für ihr tägliches Leben und soziale Netzwerke **gelten**?
RDLM	wie kann man **diese Definition** auf ihren Alltag und sozialen Netzwerken **anwenden**?
ref	wie wird **diese Definition** auf seinen Alltag und die sozialen Netzwerke **angewendet**?

subcategorisation

gelten is intransitive.
anwenden is correct in transitive construction.

(hard-to-fix error for lemma-based SMT system with inflection prediction [Fraser et al., 2012] or post-correction approach [Rosa et al., 2012]).