Machine Translation Panel

Nadir Durrani, University of Edinburgh: Operation Sequence Model

Chris Dyer, CMU: Word Classes

Spence Green, Stanford: Sparse Feature Training

Kenneth Heafield, Stanford: Huge Language Models

Stephan Peitz, RWTH Aachen: Leave One Out Training

Philip Williams, University of Edinburgh: String-to-Tree Syntax

Nadir Durrani
University of Edinburgh

Operation Sequence Model

Introduction

- A model that
  - combines benefits from Phrase-based and N-gram-based SMT
  - is based on minimal translation but memorizes like phrases
  - considers source and target contextual information across phrases
  - integrates translation and reordering into a single model

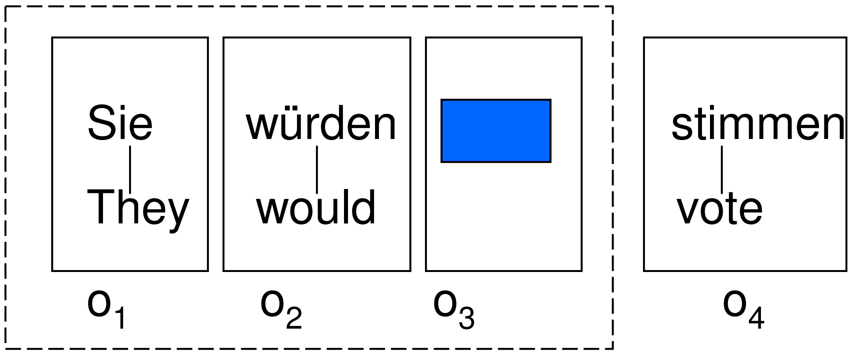
- Convert a bilingual sentence to a sequence of operations
  - Translate (Generate a minimal translation unit)
  - Reordering (Insert a gap or Jump)

- P(e,f,a) = N-gram model over resulting operation sequences

Example

Sie würden gegen Sie stimmen

They would vote against you



Context Window

Operations

- o_1 Generate (Sie, They)
- o_2 Generate (würden, would)
- o_3 Insert Gap
- o_4 Generate (stimmen, vote)
- o_5 Jump Back (1)
- o_6 Generate (gegen, against)
- o_7 Generate (Sie, you)

Model:

$$p_{osm}(F,E,A) = p(o_1,\ldots,o_N) = \Pi_i\ p(o_i | o_{i-n+1} \ldots o_{i-1})$$

How does it improve Phrase-based SMT?

- Overcomes phrasal independence assumption
  - Considers source and target contextual information across phrases
- Better reordering model
  - Translation and reordering decisions influence each
  - Handles local and long distance reorderings in a unified manner
- No spurious phrasal segmentation problem
- Average gain of +0.40 on news-test2013 across 10 pairs

Thank You !!!

Chris Dyer
CMU

Word Classes

Using Word Clusters

C10010111 → C00100001 → C11111 →

1. Cluster monolingual data
2. 500-1000 clusters
3. Use for: LMs, features

Common prefixes,
Common context

^1101100 (126689)	Merkel Obama Müller Schmidt Friedrich Steinbrück Fischer Koch Wulff Westerwelle Schäuble Schneider Löw Sarkozy Seehofer Schröder Vettel S. Putin Beck Berlusconi B. Wagner Gabriel Rösler Wolf Hoeneß Becker Weber Steinmeier Meyer Bush Clinton Bauer Schulz Schumacher Jung Romney Schäfer Klein M. Kaiser Gomez Roth K. W. Ali Hollande Gauck Heynckes
^1101101000 (2965)	Deutschland Madrid Leverkusen BSC Motors Stanley Brothers Woods bar Christus Ostdeutschland
^11011010010 (5147)	China Österreich Frankreich Russland Italien Spanien Israel Japan Großbritannien Polen Ägypten Indien Brasilien Schweden England Niedersachsen Brandenburg Portugal Irland Pakistan Australien Ungarn Kanada Belgien Nordkorea Südafrika Dänemark Thüringen Tschechien Rumänien Teheran Serbien Norwegen Südkorea Silber Kroatien Finnland Tunesien Argentinien Island Gladbach Bulgarien Bronze Osteuropa Hongkong Singapur Georgien Katar Thailand Holland
^11011010011 (4071)	Europa Griechenland Syrien Afghanistan Hessen NRW Amerika Libyen Afrika Asien Zypern Fukushima Rio Boston Kuba Gaza Mallorca Mali Wimbledon Bosnien Tibet Haiti Gorleben Sylt Palästina Hawaii Auschwitz Tschernobyl Sotschi Westeuropa Schach Lampedusa Jugoslawien Oberbayern Santiago Philadelphia Oberfranken Wembley Übersee Guantánamo Teilzeit Fort Guantanamo Ostwestfalen Tschetschenien Aufruhr Echtzeit Südeuropa Sierra Steueroasen
^1101101010 (6344)	München Frankfurt Stuttgart Köln Dortmund Hannover Bremen Düsseldorf Nürnberg Wolfsburg Augsburg Mainz Freiburg Bochum 04 Hoffenheim Mönchengladbach United Paderborn Cottbus Magdeburg Zürich Fürth Darmstadt Basel Braunschweig Lübeck Bamberg Ingolstadt Osnabrück Valencia Ried Friedrichshafen Herford Minden Koblenz Gütersloh Aalen Aue Florenz Mavericks Rottweil Schweinfurt Borken Donaueschingen Freudenstadt Calw Ravensburg Saarbrücken Reutlingen
^1101101011 (22412)	Berlin Hamburg Wien London Paris Washington Brüssel Moskau Sachsen Rom Peking Salzburg Athen Münster Leipzig Dresden Potsdam Kiel Bonn Linz Wiesbaden Bielefeld Kairo Mexiko Aachen Karlsruhe Duisburg Istanbul Mailand Tokio Damaskus Graz Luxemburg Regensburg Mannheim Kassel Rostock Flensburg Amsterdam Kalifornien Innsbruck Tripolis Würzburg Florida Jerusalem Offenbach Chicago Oberösterreich Erfurt Ulm

Using Word Clusters

C10010111 → C00100001 → C11111 →

Rule “shape” features (prefix length = 6)

X→(*dass* X *angekommen ist*, that X arrived)

C1001_X_C001101_C110100::C010111_X_C111111=1

7-gram class-based LM

Absolute discounting (d=0.5)
Separate features for transitions and emissions

	BLEU	MET	TER
Baseline	25.3	30.4	52.6
+Rule shape	25.5	30.5	52.4
+7gm LM	**26.4**	**31.0**	**51.9**

Spence Green
Stanford

Sparse Feature Training

Large-scale Discriminative Tuning

#1: **2010s ML in MT tuning**

Online convex optimization

Arbitrary, overlapping features

#2: **Large tuning sets**

Fast decoding and updating

Bitext tuning...

See our poster and talk for details

WMT14 Shared Task Results

Uncased BLEU results

	dense−dev	**features−dev**	**2014 rank**
Fr−En	19.6	20.0	1
En−De	32.0	32.5	1

Tune: 13.5k sentences (2008−2012)

Models have 200−300k features

Kenneth Heafield
Stanford

Huge Language Models

Impact of Big Language Models

Target	**Base Rank**		**+LM Rank**	**ΔBLEU**
Czech	5–6	$\rightarrow$	1–3	+0.6
Hindi	4–5	$\rightarrow$	3	+1.4
Russian	6–7	$\rightarrow$	4–5	+1.2
German	8–10	$\rightarrow$	3–6	+0.5

After the evaluation: Hindi–English +0.9 BLEU

Download multiple LMs and training data from
`statmt.org/ngrams`

English: 1.8 trillion tokens

n	Unique *n*-grams
1	2,640,258,088
2	15,297,753,348
3	61,858,786,129
4	156,775,272,110
5	263,690,452,834

Current work: approximate LM storage.

Stephan Peitz
RWTH Aachen

Leave One Out Training

Consistent Phrase Training

State of the art

- **Heuristic extraction of phrases using word alignments**
- **Compute translation probabilities as relative frequencies**

Issues of this heuristic

- **Extract from likely alignment?**
- **Models used in decoding are not considered $\Rightarrow$ inconsistency**

Forced decoding

- **Run decoder on training data**
- **Count used phrases, recompute probabilities**
- **Apply leave-one-out to counteract overfitting**

Leave-One-Out

- Occurrences of a phrase in a sentence pair (f_n, e_n) are subtracted from the phrase counts obtained from the full training data

$$p_{l1o,n}(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f},\tilde{e}) - C_n(\tilde{f},\tilde{e})}{\sum_{\tilde{f}'} C(\tilde{f}',\tilde{e}) - C_n(\tilde{f}',\tilde{e})}$$

- Singleton phrases get a low probability

Consistent Phrase Training using Leave-One-Out

- **Publications**
  - Phrase-based [Wuebker & Mauser+ 10, Wuebker & Ney 13]
  - Hierarchical [Peitz & Mauser+ 12, Peitz & Vilar+ 14]
- **Improvements: 0.5-1.5 BLEU**
- **Reducing phrase-table size to 5-20% of the original size**
- **Systems using phrase/rule training:**
  - WMT 2011 (RWTH, German→English, phrase-based)
  - IWSLT 2011 (RWTH, German→English, phrase-based)
  - IWSLT 2012 (RWTH, German→English, hierarchical)
  - BOLT 2012 (RWTH, Chinese→English, hierarchical)
  - OpenMT 2012 (NRC, Chinese→English, phrase-based)
- **Implemented in RWTH's translation toolkit Jane**
  `http://www.hltpr.rwth-aachen.de/jane`

References

[Peitz & Mauser$^+$ 12] S. Peitz, A. Mauser, J. Wuebker, H. Ney: Forced Derivations for Hierarchical Machine Translation. In *International Conference on Computational Linguistics*, pp. 933–942, Mumbai, India, Dec. 2012. 4

[Peitz & Vilar$^+$ 14] S. Peitz, D. Vilar, H. Ney: Simple and Effective Approach for Consistent Training of Hierarchical Phrase-based Translation Models. In *Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April 2014. 4

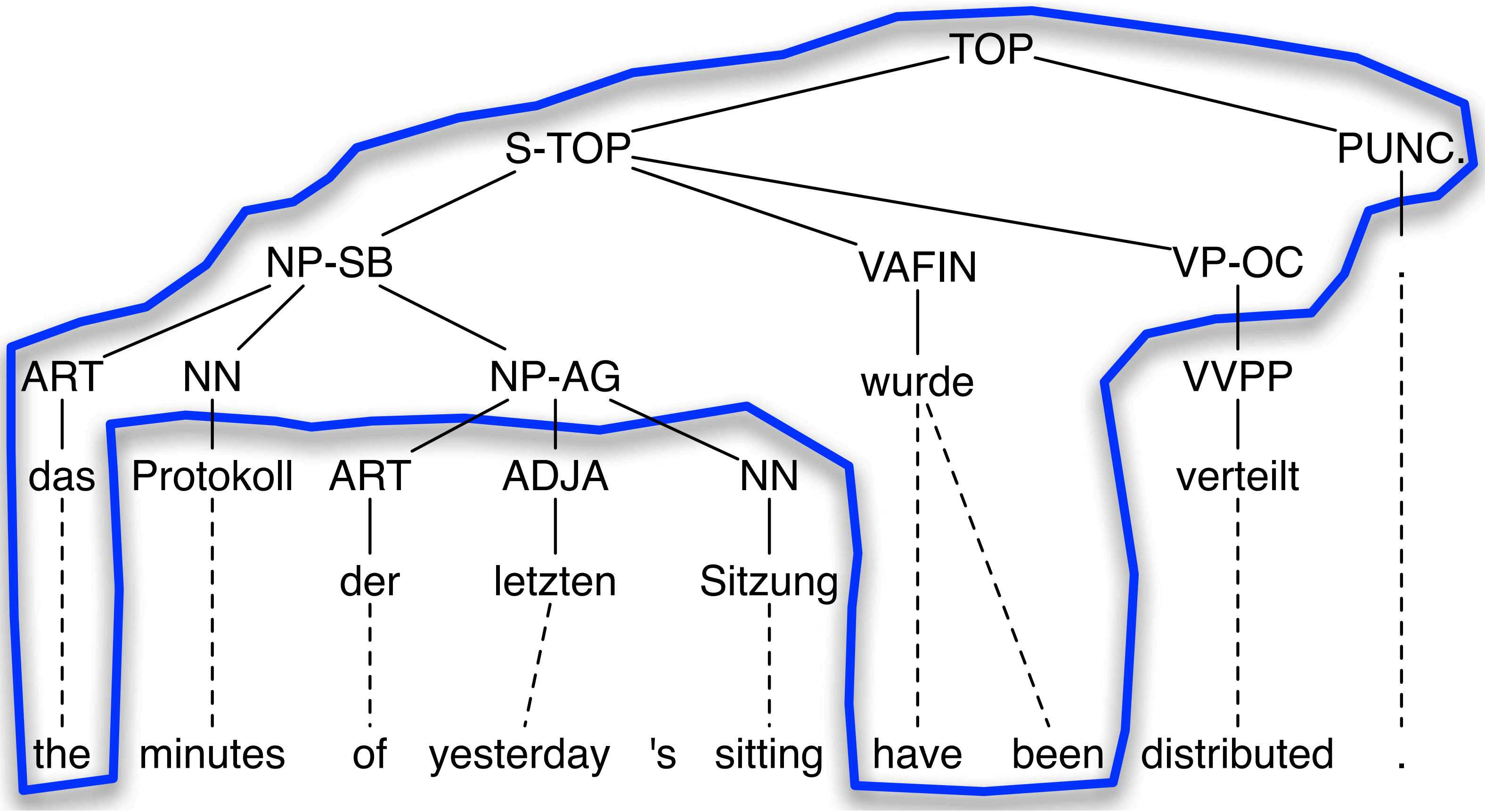
[Wuebker & Mauser$^+$ 10] J. Wuebker, A. Mauser, H. Ney: Training Phrase Translation Models with Leaving-One-Out. In *Annual Meeting of the Assoc. for Computational Linguistics*, pp. 475–484, Uppsala, Sweden, July 2010. 4

[Wuebker & Ney 13] J. Wuebker, H. Ney: Length-incremental Phrase Training for SMT. In *ACL 2013 Eighth Workshop on Statistical Machine Translation*, pp. 309–319, Sofia, Bulgaria, Aug. 2013. 4

Philip Williams
University of Edinburgh

String-to-Tree Syntax

uedin-syntax: string-to-tree



TOP $\rightarrow$ the X_1 X_2 have been X_3 X_4 | das NN_1 NP-AG$_2$ wurde VP-OC$_3$ PUNC.$_4$

uedin-syntax: string-to-tree extensions

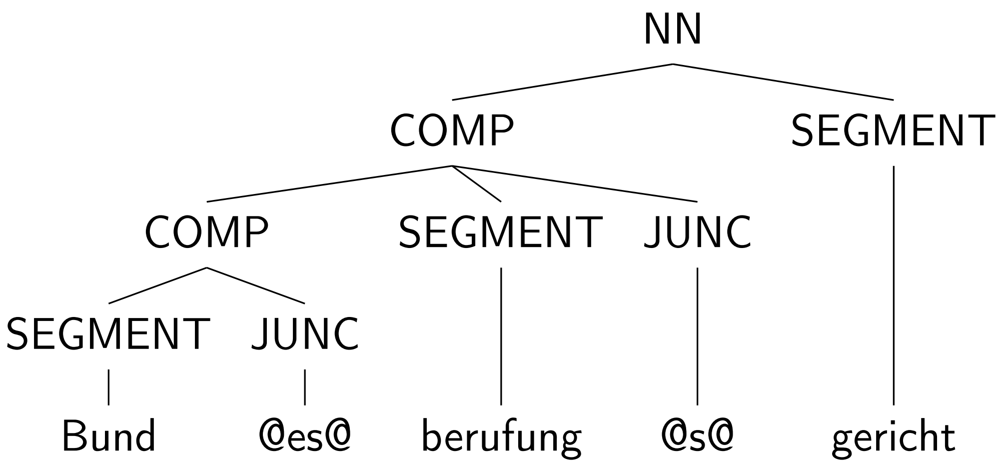
Use syntactic structure to help model other aspects of target-side grammar.

Example 1. Agreement

$$\text{TOP} \rightarrow \text{the } X_1\ X_2 \text{ have been } X_3\ X_4 \ \mid \text{das } \text{NN}_1\ \text{NP-AG}_2 \text{ wurde } \text{VP-OC}_3\ \text{PUNC.}_4$$

$$\langle \text{NN}_1\ \textsc{agr} \rangle = \langle \text{wurde}\ \textsc{agr} \rangle$$

Example 2. Compound Splitting



Best constrained system for:

- English-German
- German-English
- Hindi-English (tied with CMU)