Noun noun compound analysis

Lonneke van der Plas

18 March 2013

SFB Klausurtagung
Universität Stuttgart
Noun noun compounds

chocolate cake = a cake made of chocolate
Christmas cake = a cake served at Christmas
Wind erosion= erosion by wind (agent)
taxi driver = driver of a taxi (theme)

Sequences of nouns functioning as a single noun (Downing, 1977)

Covert relation between the two nouns
Translation of noun noun compounds

apple pie = a cake made of apples → tarte aux pommes
Christmas cake = a cake served at Christmas → gateau de Noël
wind erosion = erosion by wind (agent) → érosion par le vent
taxi driver = driver of a taxi (theme) → chauffeur de taxi

- Machine translation would benefit from compound analysis
  → Fabienne Cap, Alexander Fraser (2010)
  How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing
  → Alexander Fraser et al. (2012)
  Modeling Inflection and Word-Formation in SMT (SFB 732:D2)
Inventories or unrestricted lists of relations?

- In linguistic theory and computational linguistics either
  1. inventory-based → restricted set of semantic relations
     Christmas cake = cake TEMPORALLY-LOCATED-OBJECT Christmas
  2. integrational → no restriction on relations
     Christmas cake = a cake served at Christmas

- Criticism@1: inventories cannot capture variety in relations, are too vague, are impoverished representations
- Criticism@2: Little explanatory power

Task-based evaluation:

- Query reformulation: unrestricted paraphrases
- MT: inventories that correlate with classes of compound translations
Cross-lingual approach to noun noun compound analysis

- Use correlation between semantics and translations of noun compounds between several languages
  - to determine the level of specificity in representation (unlimited variety of paraphrases - restricted relations in inventory)
  - to create universal inventory
- Machine translation as task-based evaluation
- Automatic extraction of noun noun compounds from parallel corpora
Relation to my previous work

- **Distributional semantics** (PhD work)
  
  *Finding semantically related words by comparing their contexts*

  Generalize across noun clusters:
  carrot cake, apple cake vs Christmas cake, wedding cake

- **Semantic role labelling** (post-doc work)
  
  *Detection of semantic arguments associated with predicate*

  Apply semantic role labelling on deverbal nouns:
  Wind erosion (agent) vs taxi driver (theme)

- **Cross-lingual and multi-lingual approaches** (PhD and post-doc work)

  Working on several languages, using parallel corpora, tools from MT
Prior work in Computational Linguistics

- **Inventory-based**: Restricted set of semantic relations
  - Prepositional paraphrasing (Lauer, 1995; Lapata and Keller, 2004)
  - General semantic classes (Nastase and Szpakowicz, 2003; Girju et al., 2005; Semeval task4, 2007, O Séaghdha, 2008; Tratz and Hovy, 2010;)
  - Suitable for statistical classification

- **Integrational**: no restriction on relations
  - Structural representations: knowledge-rich approaches for closed domain (Finin, 1980) emergent representations (Rosario, 2002)
  - Paraphrasing verbs (Nakov and Hearst, 2006; SemEval-2 Task 9, Butnario et al., 2010; Dobó and Pulmann, 2011)

- Crosslingual features to do better automatic classification, inventory-based (Girju, 2007)
Embedding in Stuttgart

- Covert events
  - SFB 732:D6: Lexical-semantic factors in event interpretation, Sabine Schulte im Walde, Sebastian Padó et al.
- Romance languages, conversion
  - SFB 732:B7: Conversion in French and Italian, Daniela Marzo et al.
  - Achim Stein
- Alignment of compounds in parallel corpora
  - Fabienne Cap, Alex Fraser (2010) How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing
  - Marion Weller and Ulrich Heid (2012): Analyzing and Aligning German compound nouns
  - SFB 732:D2 → Alex Fraser et al. (2012). Modeling Inflection and Word-Formation in SMT.
- Determine non-compositionality in compounds
  - SemRel project, Sabine Schulte im Walde: Modelling the compositionality of German multi-word expressions
Collaborations outside Stuttgart

- University of Tilburg (Menno van Zaanen, Walter Daelemans)
  Project: Automatic Compound Processing (AuCoPro)
- University of Geneva (Pierrette Bouillon)
  Project: Annotation of noun noun compounds in parallel corpora
Interaction across modules: syntactic, shallow-semantic, lexico-semantic modules, ontological knowledge. Parallel: distributional approaches + hand-built lexicons, comprehension + production (MT) → take direction of translation into account

Non-canonical data: weakly supervised, using clusters of nouns to circumvent data bottleneck, focus on languages other than English, (oral) parallel corpora

Search: MT: translation into languages with higher levels of specification as a testbed
Thank you!
Prior work in Linguistics

- **Inventory-based**: restricted set of semantic relations
  - Generalised verbs (Lees, 1970),
  - Recoverable Deletable Predicates (Levi, 1978)
  - Criticism: inventories cannot capture variety in relations, are too vague, are impoverished representations

- **Pro-verb**: underspecified representation in semantics, leave analysis to pragmatics

- **Integrational**: no restriction on relations
  - blending (Ryder, 1994; Coulson, 2001)
  - event frames (Downing, 1977; Coulson, 2001)
  - frame semantics (Johnston and Busa, 1996)
Debates about suitable classification for compounds. Proposal by Bisetto and Scalise (2009) based on cross-linguistic research.

Compounds:
- Subordinate:
  - Endo: steamboat, love story
  - Exo: pickpocket, loudmouth
- Attributive:
  - Endo: blackbird, blackboard
  - Exo: green-eyed, greybeard
- Coordinate:
  - Endo: bittersweet, girlfriend
  - Exo: mother-child