

# Context as Linguistic Bridges

Maarten van Gompel



## Context

- **Intuïtie:** “*Context speelt een belangrijke rol in taal*”
- Betekenis van woorden, zinsneden, zinnen etc... is afhankelijk van de context waarin ze verschijnen.
- Context helpt eventuele ambiguïteit in betekenis op te lossen

## Voorbeeld: een woord

### “bank”

Volgens van Dale:

- ① zitmeubel voor meer dan één persoon
- ② verkooptafel: toonbank
- ③ werktafel: draaibank
- ④ door de natuur gevormde ondiepte
- ⑤ instelling die gelden beheert en uitleent
- ⑥ (bij kansspelen) inzet van de hoofdspeler tegen alle andere samen
- ⑦ centrale bewaar- en uitwisselingsplaats; = opslagplaats: bloedbank, vacaturebank

## Voorbeeld 1: een woord in context

### “bank”

- “De **bank** heeft al mijn geld verloren.”
- “De hond mag op de **bank** liggen.”

## Word Sense Disambiguation

- Context helpt bij het vinden van de betekenis
- Het automatisch disambiguëren van de juiste betekenis van een woord heet **Word Sense Disambiguation**.

## Betekenis en Vertalingen

- Het vinden van de juiste betekenis is belangrijk bij vertalen
- Een woord kan anders vertalen naar gelang de betekenis

## Voorbeeld 1: Een woord in context

### “bank”

- “*De bank heeft al mijn geld verloren.*”
  - “*La banque à perdu tout mon argent.*”
- “*De hond mag op de bank liggen.*”
  - “*Le chien peut se mettre sur le canapé.*”

## Voorbeeld 2: Een woord in context

**“tempo”** (portugees)

- “*Faz bom **tempo** hoje.*“
  - “*Het is lekker **weer** vandaag*“
- “*Ele tem **tempo** para mim.*“
  - “*Hij heeft **tijd** voor mij.*“

## Onderzoek

- De rol van contextinformatie in automatische vertaling
- **Hoofdvraag:** In hoeverre kunnen we automatische vertaling verbeteren door contextinformatie uit de bronstaal mee te nemen?
- Empirische studie: Schrijven van software, trainen op trainingsdata, testen op testdata, vergelijken met anderen, conclusies afleiden

## Aanpak WSD

- We bouwen een **classifier-gebaseerd** cross-lingual WSD systeem in navolging van eerdere studies (Hoste et al., 2002; Hendrickx et al., 2002)
  - **memory-based learning / TiMBL software / IB1 algoritme**
- Het systeem leert op basis van voorbeelden hoe een woord met bepaalde **contextuele kenmerken** vertaald wordt.
  - Lokale contextinformatie (woordvorm):  $x$  woorden links,  $y$  woorden rechts
  - Linguïstisch verrijkte lokale contextinformatie (PoS+lemma)
  - Globale contextinformatie (woordvorm) (binary bag of word features)
- Deelname aan drie SemEval taken
  - SemEval 2010: Cross Lingual Lexical Substitution (Engels naar Spaans) (Mihalcea et al., 2010)
  - SemEval 2010: Cross Lingual Word Sense Disambiguation (Engels naar Nederlands/Spaans) (Lefever and Hoste, 2010)
  - SemEval 2013: Cross Lingual Word Sense Disambiguation (Engels naar Nederlands/Spaans/Frans/Duits/Italiaans) (Lefever and Hoste, 2013)

## Deelvragen & Resultaten

- ① Welke informatie draagt het meeste bij tot een juiste vertaling?
  - Lingüistisch-geïnformeerd of niet?
    - Lemmas features helpen, PoS juist niet
  - Lokale en/of globale features?
    - Globale features vallen tegen ondanks wat eerdere succesen
  - Contextgrootte?
    - Kleine contextgrootte werkt het best (1 links, 1 rechts)
- ② Welke optimalisatietechnieken kunnen we toepassen?
  - Arbiter Voting (WSD1, 2010)
  - Automatische feature selectie (WSD2, 2013)
  - Classifier parameteroptimalisatie (WSD2, 2013)

## Competitie Resultaten

- Winnende scores in de Cross Lingual WSD taak (2010), later ook een aantal winnende scores in 2013

## Een nieuwe insteek

- ① kunnen we fragmenten automatisch vertalen met een **anderstalige context**?
  - het gebruiken van korte fragmenten van de ene taal in een anderstalige context (bv. zin)
- ② **codewisseling/code switching:** het overspringen tussen twee talen

## Voorbeeld 1: Codewisseling met één woord

L1 = Nederlands, L2 = Duits

- ① "Alles was er sagt ist **altijd** falsch"
- ② "Alles was er sagt ist **immer** falsch"

## Voorbeeld 2: Codewisseling met langer fragment

L1 = Nederlands, L2 = Duits

- ① "Alles **wat hij zegt** ist immer falsch"
- ② "Alles **was er sagt** ist immer falsch"

## Opzet

- Zit er waarde in dit idee?
  - Ja, denk aan vertaalhulp systemen
- Kunnen we dit net als voorheen met classifiers oplossen? → **pilot study** (hfd 4)
  - Ja
  - Maar; we hebben geen echte representatieve data om op te testen
- Testdata samengesteld en daarmee een nieuwe SemEval taak organiseren (hfd 5)
  - 4 taalparen (L1-L2): Engels-Spaans, Engels-Duits, Frans-Engels, Nederlands-Engels
  - Handmatige dataverzameling uit verschillende bronnen
  - 6 kandidaten doen mee met onze taak
- Met de nieuwe testdata ons eigen systeem (*colibrita*) opnieuw testen en verbeteren (hfd 6)

## Van WSD naar SMT

### Van Word Sense Disambiguation naar (phrase-based) Statistical Machine Translation

- Naar mate we langere fragmenten bezien begeven we ons steeds meer op de het terrein van de SMT
- Kunnen SMT technieken ook toegepast worden in het code-switching scenario?
  - Ja, twee deelnemers aan onze taak laten dit duidelijk zien
  - Resultaten beter dan pure classifier aanpak
- **Beste van beide werelden?** Kunnen we de expliciete modellering van de classifiers integreren in SMT en heeft dat meerwaarde? (hfd 6,7)
  - We vergelijken classifiers, classifiers geïntegreerd in SMT, en pure SMT

## Integratie van classifiers in SMT

- Op welke manieren kunnen we classifiers in SMT integreren?
  - (*classifiertype*: classifier experts vs. monolithisch) , (*weegmethode*: replace vs append)
  - Minimale verschillen met weinig effect in het algemeen, geen eenduidige conclusie te vormen
- Heeft data zonder linguïstische verrijking een meerwaarde in deze opstelling?
  - Nee, de **hoofdconclusie** is dat we explicet proberen te modelleren wat de bestaande modellen (vertaalmodel + taalmodel) implicit al goed genoeg dekken. Daar waar contextinformatie uit de bronstaal een disambigueringse rol kan spelen, is hij al potentieel beschikbaar in het vertaalmodel.

## Conclusies

- ① Contextinformatie is nuttig voor vertaling (ook met anderstalige context)
- ② We sluiten een lijn van onderzoek (Stroppa et al., 2007; Haque et al., 2011) (de integratie van niet-linguistische-geinformeerde classifier-gebaseerde WSD technieken in (PB)SMT) en concluderen dat deze integratie geen meerwaarde biedt.
- ③ We creëren vaak marginale lokale verschillen en concluderen dat een gedegen evaluatie op automatische metrieken en bij een gebrek aan meerdere referentieverhalingen erg lastig is.
- ④ We benadrukken het belang van testen op meerdere datasets, meerdere taalparen, en het doen van significantietests om niet te snel conclusies te trekken.

## References |

- Haque, R., Naskar, S., van den Bosch, A., and Way, A. (2011). Integrating source-language context into phrase-based statistical machine translation. *Machine Translation*, 25(3):239–285.
- Hendrickx, I., Van den Bosch, A., Hoste, V., and Daelemans, W. (2002). Dutch word sense disambiguation: Optimizing the localness of context. In *Proceedings of the Workshop on word sense disambiguation: Recent successes and future directions*, pages 61–65, Philadelphia, PA.
- Hoste, V., Hendrickx, I., Daelemans, W., and Van den Bosch, A. (2002). Parameter optimization for machine learning of word sense disambiguation. *Natural Language Engineering*, 8(4):311–325.
- Lefever, E. and Hoste, V. (2010). Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 15–20, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Lefever, E. and Hoste, V. (2013). SemEval-2013 Task 10: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics*.
- Mihalcea, R., Sinha, R., and McCarthy, D. (2010). Semeval 2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.
- Stroppa, N., van den Bosch, A., and Way, A. (2007). Exploiting source similarity for SMT using context-informed features. In Way, A. and Gawronska, B., editors, *Proceedings of the 11th International Conference on Theoretical Issues in Machine Translation (TMI 2007)*, pages 231–240, Skövde, Sweden.