

Supersense Tagging for Arabic: The MT-in-the-Middle Attack



Nathan Schneider

Behrang Mohit

Chris Dyer

Kemal Oflazer

Noah A. Smith

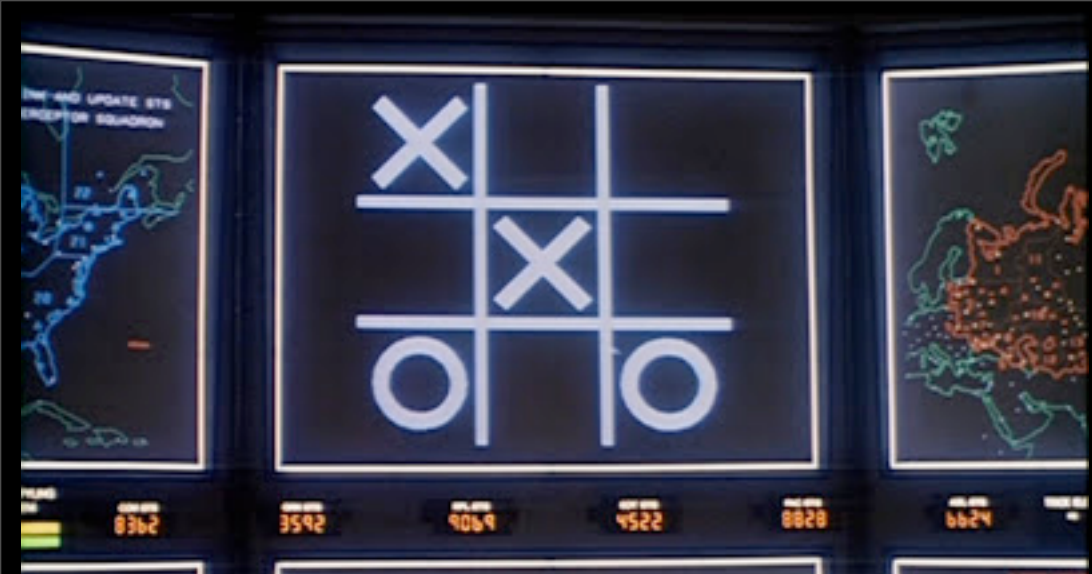


Carnegie Mellon

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar



الصندوق الوطني للبحوث العلمية
Qatar National Research Fund



GAMEPLAN

Supersense Tagging

Baselines

MT-in-the-Middle

Analysis

Outlook

Supersense Tagging



- A coarse form of **word sense disambiguation** (partitioning of WordNet synsets)
- Generalizes NER beyond proper names;
26 noun categories (Ciaramita & Johnson 2003)

SOCIAL

Pierre Vinken, 61 years old, will join the board as a nonexecutive director

PERSON TIME SOCIAL GROUP PERSON

- Categories broadly applicable across domains
- Scheme suitable for direct annotation (Schneider et al. 2012)

Supersense Tagging



- **English** resources
 - ▶ WordNet (Fellbaum 1998)
 - ▶ Tagger trained on English SemCor (Ciaramita & Altun 2006) **77% F₁ in-domain**
- **Arabic** resources
 - ▶ Arabic WordNet (El Kateb et al. 2006)
 - ▶ Named entities in OntoNotes (Hovy et al. 2006)
 - ▶ Supersense-tagged Wikipedia corpus (Schneider et al. 2012)
65k words—1/6 the size of SemCor

Baselines



- **Heuristic matching** of Arabic WordNet entries + OntoNotes NEs
 - ▶ only covers 33% of nouns in our corpus

	P	R	F ₁
Ann-A	32	16	21.6
Ann-B	29	15	19.4

- **Unsupervised sequence model**
 - ▶ feature-rich (Berg-Kirkpatrick et al. 2010)

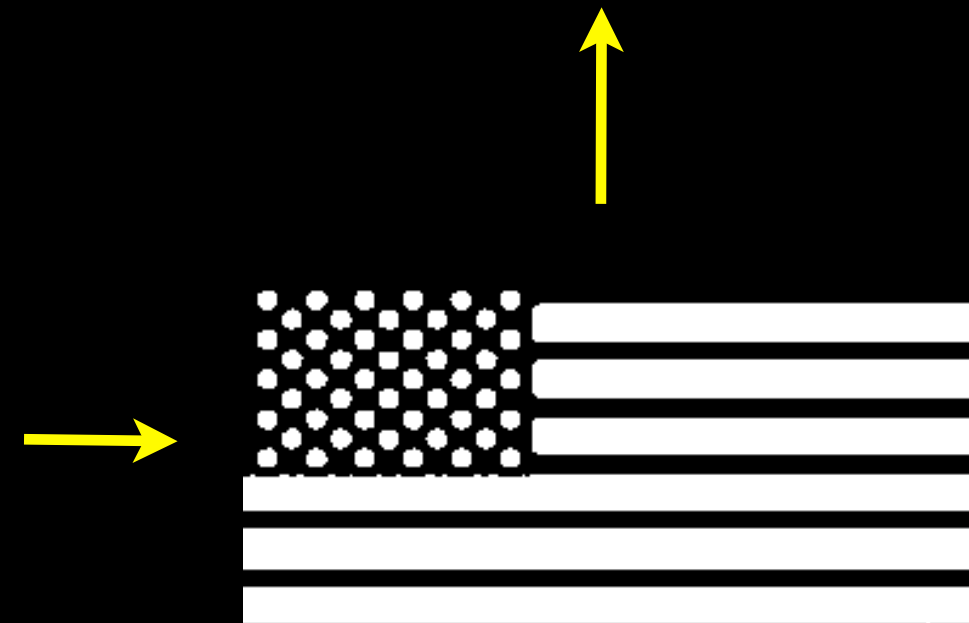
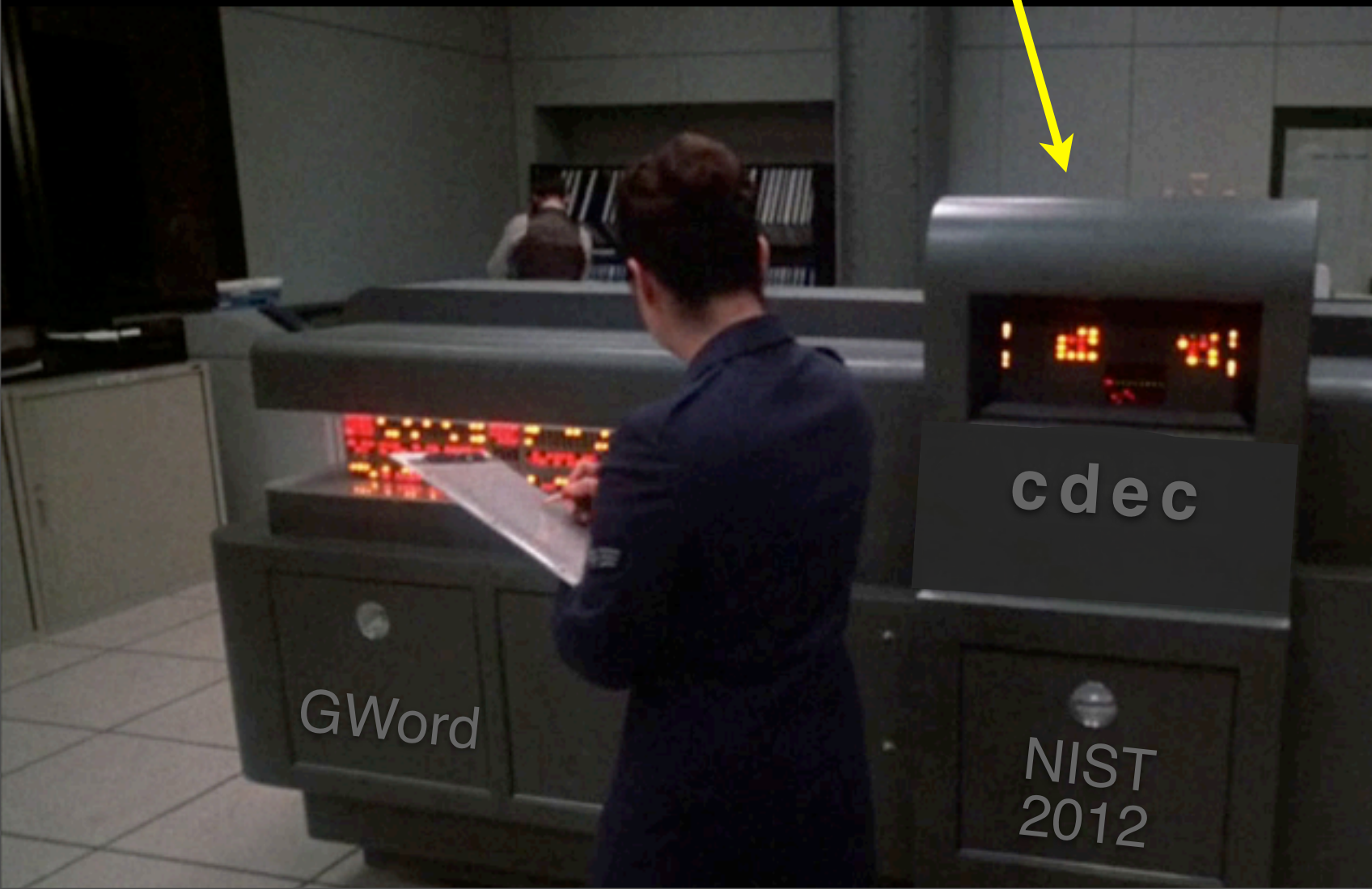
	P	R	F ₁
Ann-A	20	16	17.5
Ann-B	14	10	11.6

[evaluating on Arabic Wikipedia test set—
18 articles, 40k words]

MT-in-the-Middle

(cf. Zitouni & Florian 2008;
Rahman & Ng 2012)

تتكون الذرة من سحابة من الشحنات السالبة (الإلكترونات)
تحوم حول نواة موجبة الشحنة صغيرة جدا في الوسط .



MT-in-the-Middle



تتكون الذرة من سحابة من الشحنات السالبة (الإلكترونات)

تحوم حول نواة موجبة الشحنة صغيرة جدا في الوسط .

The **corn** is composed of negative **shipments** (**electronics**)

PLANT

ARTIFACT COGNITION

cloud hovering over the **nucleus** of a very small positive

BODY

shipment in the **center** .

ARTIFACT

LOCATION

MT-in-the-Middle



تتكون الذرة من سحابة من الشحنات السالبة (الإلكترونات)

تقوم حول نواة موجبة الشحنة صغيرة جدا في الوسط .

The **corn** is composed of negative **shipments** (**electronics**)

PLANT

ARTIFACT COGNITION

cloud hovering over the **nucleus** of a very small positive

BODY

shipment in the **center** .

ARTIFACT

LOCATION

MT-in-the-Middle



تتكون الذرة من سحابة من الشحنات السالبة (الإلكترونات)
COGNITION ARTIFACT PLANT
تقوم حول نواة موجبة الشحنة صغيرة جدا في الوسط .

The corn is composed of negative shipments (electronics)

cloud hovering over the nucleus of a very small positive
BODY

shipment in the center .
ARTIFACT LOCATION

MT-in-the-Middle



- Heuristic lexicon matching:

	P	R	F ₁
Ann-A	32	16	21.6
Ann-B	29	15	19.4

- MT-in-the-Middle:

	P	R	F ₁
Ann-A	37	31	33.8
Ann-B	38	32	34.6

MT-in-the-Middle



- MT-in-the-Middle:

- Hybrid:

	P	R	F ₁
Ann-A	37	31	33.8
Ann-B	38	32	34.6

	P	R	F ₁
Ann-A	35	36	35.5
Ann-B	36	36	36.0

Analysis



- Pipeline has many places for noise: MT, English supersense tagging, and projection
- We focus on the impact of translation

Analysis



- Compare **cdec** vs. an off-the-shelf Arabic-English system from **QCRI**
- Translation quality:

	BLEU	METEOR	TER
QCRI	32.86	32.10	0.46
cdec	28.84	31.38	0.49

- ...but for MTiTM supersense tagging, **cdec** is consistently better (by 2–4 points). Why?

Analysis



- Observation: overall MT scores do not necessarily measure *preservation of coarse lexical semantics*
 - ▶ We really care about (rough) semantic adequacy for noun phrases
 - ▶ We elicited **lexical translation acceptability** judgments for a sample of sentences (cf. Carpuat 2013: SSSST)

Analysis



- Lexical acceptability rates: 91.9% for **QCRI**, 90.0% for **cdec**
- Example errors
 - ▶ *corn, maize* for *atom*
 - ▶ *shipments* for *charges*
 - ▶ *electronics* for *electrons*
 - ▶ transliteration: *IMAX* for *EMACS*,
genoa lynx for *GNU Linux*

Analysis



- So lexical translation is mostly OK, and **QCRI** does slightly better at it
- **cdec**'s strength: providing better input to projection
 - ▶ It produces *word* alignments, whereas **QCRI** gives *phrase* alignments

Outlook



- Supersense tagging can be accomplished (noisily) for a language so long as it can be automatically translated to English
- Further gains should come from:
 - ▶ better MT—lexical translations and word alignments
 - ▶ better English supersense tagging
 - ▶ better lexicon & corpus resources

Thanks

- Francisco Guzman & Preslav Nakov @ QCRI
- Wajdi Zaghouani
- Waleed Ammar
- QNRF
- All of you for listening!