Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop

Alfredo Alba, Anni Coden, **Anna Lisa Gentile**,
Daniel Gruhl, Petar Ristoski, Steve Welch

IBM Research

Motivation

sushirritto

Animal style fries

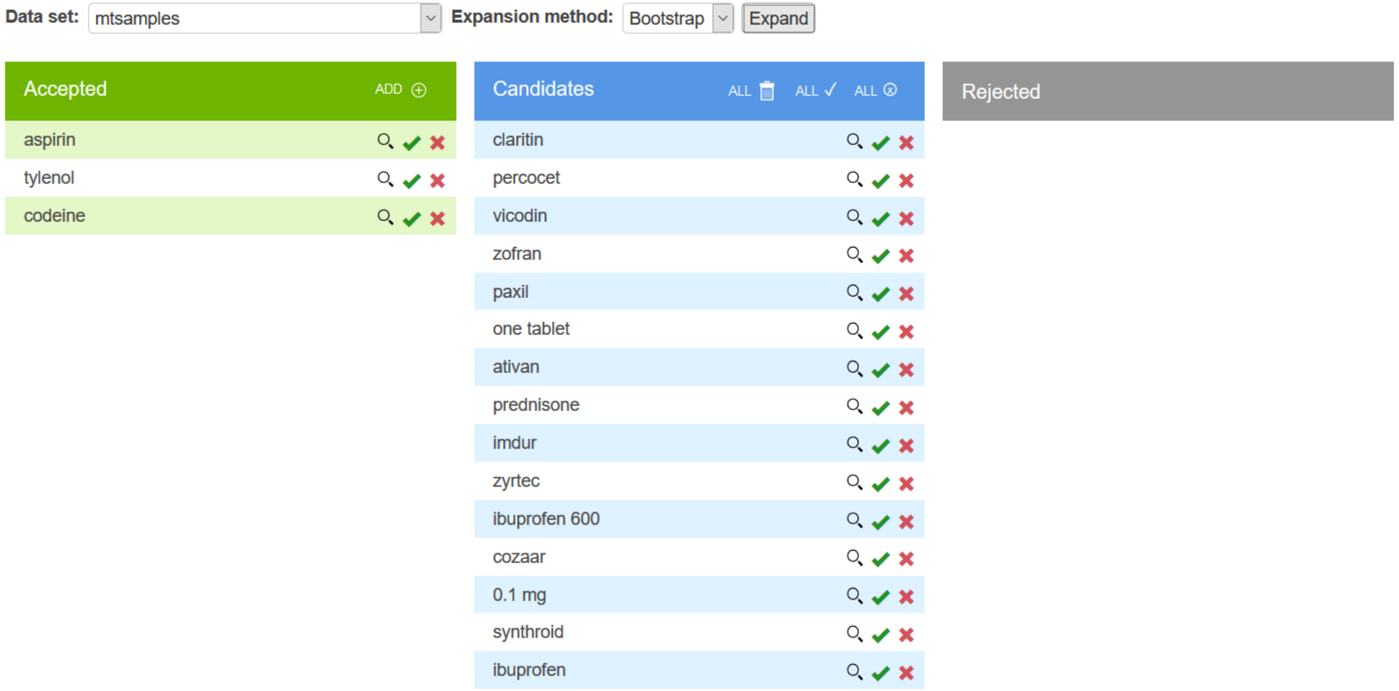
- 1 tablespoon plus 1 teaspoon brown (or white) sugar
- 1½ teaspoon salt
- 2½ tablespoons sesame oil
- vegetable oil

Motivation

- extract information from a **novel** corpus
- what are the relevant **concepts** in the domain?
- limited **domain** and **language** knowledge
- **IDEA**: combine **statistical** techniques with **user-in-the-loop**

Domain Learning Assistant

- Start with a small number of seeds (1)
- Get suggestions of new surface forms
- The user accept/reject



Data set: mtsamples

Expansion method: Bootstrap

Accepted

- aspirin
- tylenol
- codeine

Candidates

- claritin
- percocet
- vicodin
- zofran
- paxil
- one tablet
- ativan
- prednisone
- imdur
- zyrtec
- ibuprofen 600
- cozaar
- 0.1 mg
- synthroid
- ibuprofen

Rejected

Finding **concept** candidates

EN The safety and efficacy of **filgrastim** are similar in adults and children receiving cytotoxic chemotherapy

ES La eficacia y la seguridad del **filgrastim** son similares en los adultos y en los niños tratados con quimioterapia citotóxica

IT La sicurezza e l'efficacia del **filgrastim** sono simili negli adulti e nei bambini sottoposti a chemioterapia citotossica

DE Die Wirksamkeit und Unbedenklichkeit von **Filgrastim** ist bei Erwachsenen und bei Kindern , die eine zytotoxische Chemotherapie erhalten , vergleichbar

Finding **concept** candidates

EN Plasma elimination half-life of oral **pravastatin** is 1.5 to 2 hours.

IT L’emivita plasmatica di eliminazione del **pravastatin** orale é compresa tra un’ora e mezzo e due ore.

Finding **concept** candidates

Candidates: {eggs, flour}

“mix **eggs** and **flour**” → mix <candidate> and <candidate>

mix <candidate> and <candidate> → “mix **sugar** and **butter**”

Candidates: {eggs, flour, sugar, butter}

“melt the **butter**” → melt the <candidate>

…

Finding **concept** candidates

Candidates: {uova, farina}

“amalgamare **uova** e **farina**” → amalgamare <candidate> e <candidate>

amalgamare <candidate> e <candidate> → “amalgamare **zucchero** e **burro**”

Candidates: {uova, farina, zucchero, burro}

“sciogliere il **burro**” → sciogliere il <candidate>

…

Multi-lingual experiment

HYPOTHESIS: same behavior, regardless of the language

- we start with very **few seeds** (one could be sufficient) for **each language**
- we extract context patterns and use them to **generate new candidates**
- we ask to **user** to **accept/reject** the candidates
- we repeat for a fixed number of iterations in all languages

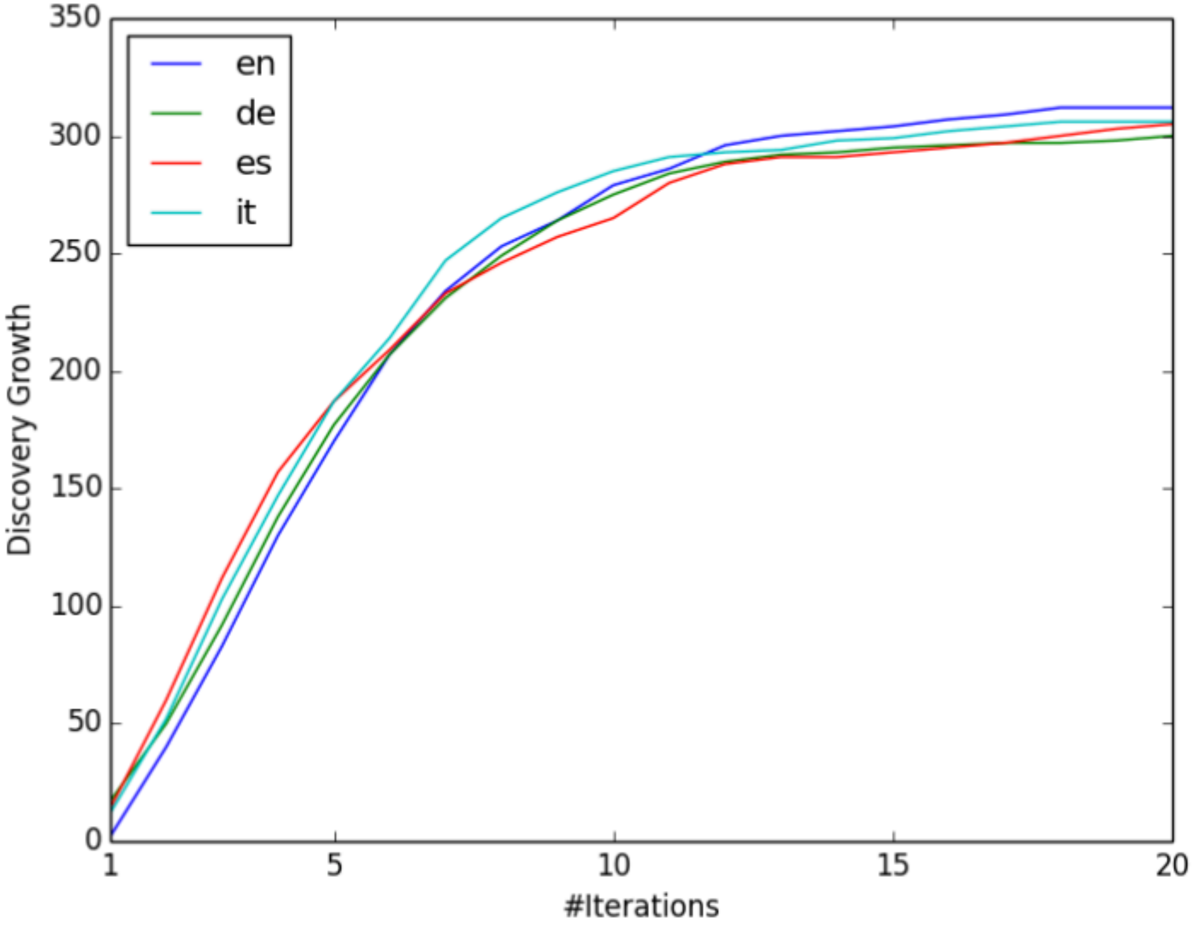
Multi-lingual experiment: **Drug Discovery**

- **DATA**: parallel corpus from the European Medicines Agency (EMEA)
  - documents related to **medicinal products**
  - translations into 22 official languages of the European Union
  - 1,500 documents for most of the languages
  - we used 4 languages (**en**, **es**, **it**, **de**)
- **TASK**: build a lexicon of clinical drugs
- **user-in-the-loop** simulated by constructing a Gold Standard (GS) of drugs names extracted from Linked Open Data (we used **DBpedia** http://dbpedia.org)

IT ES DE EN

Drug Discovery: One seed

- **initial seeds:** single seed
  - One drug name which appears in each corpus (e.g. “irbesartan”)
- 20 iterations
- learning curves for all languages are comparable

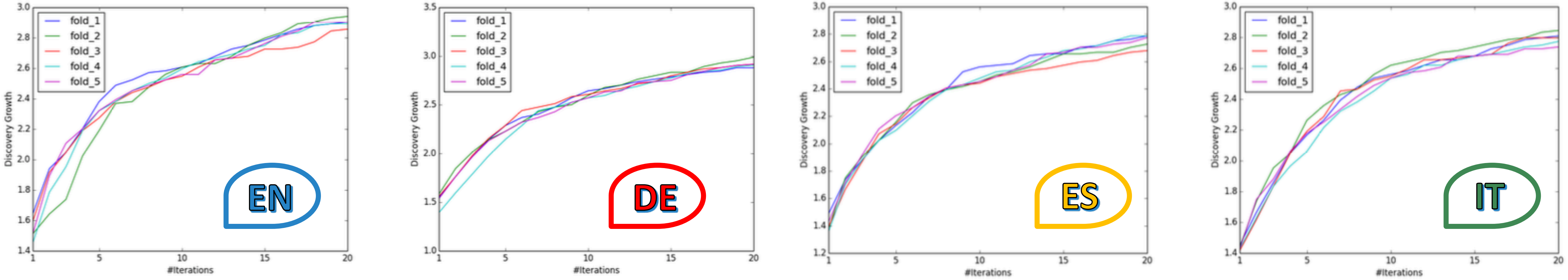


Discovery growth for glimpse for English (en), Italian (it), Spanish (es) and German (de). Average correlation amongst all languages r = 0.998.

IT ES DE EN

Drug Discovery: Linked Data seeds

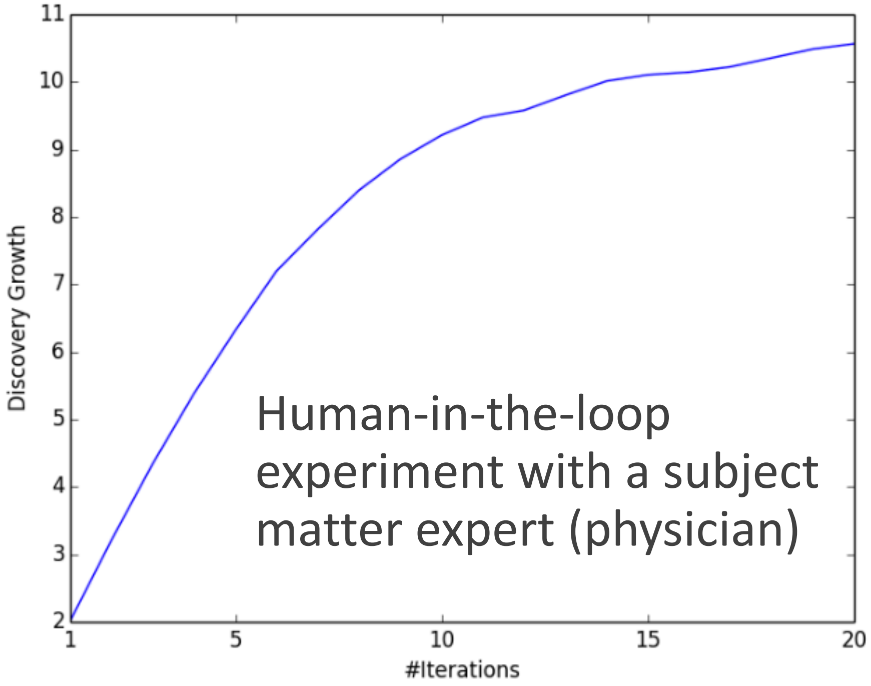
- **initial seeds:** 20% of available Linked Data (DBpedia)
  - 5-fold validation (randomly selected 20%, same drugs for all languages)
  - choice of initial seeds **does not impacts** the results



(a) *EMEA* English ($r = 0.991$) **(b)** *EMEA* German (r= 0.995) **(c)** *EMEA* Spanish ($r = 0.994$) **(d)** *EMEA* Italian ($r = 0.996$)

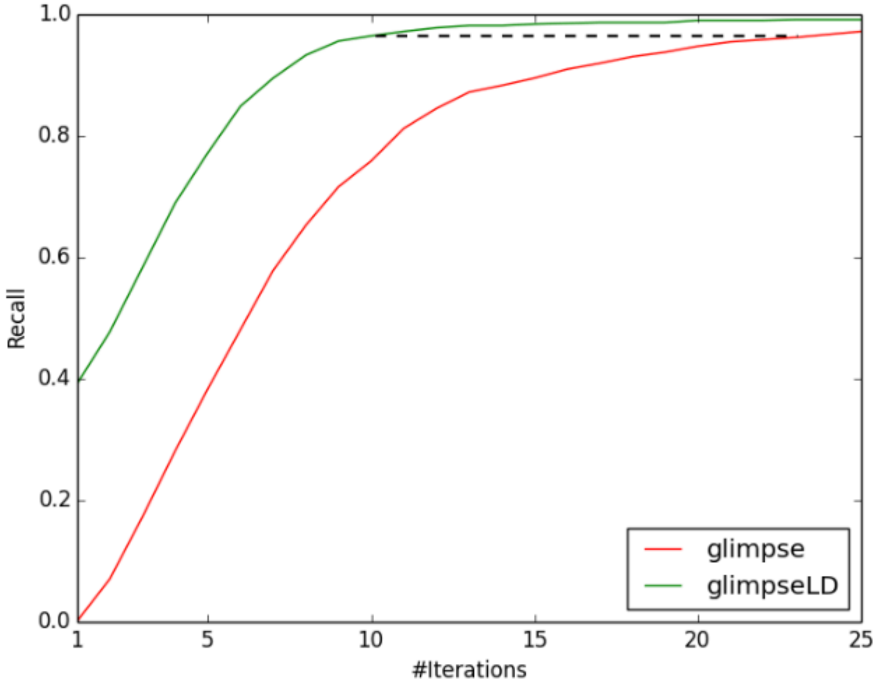
Discovery growth with 5-fold cross validation on the EMEA dataset using DBpedia as seeds. Each plot shows the discovery growth for each of the randomly generated 5 folds and reports the Pearson correlation (r) amongst them.

Drug Discovery: benefit of Linked Data



Human-in-the-loop experiment with a subject matter expert (physician)

(a) *Discovery growth for glimpseLD.*



(b) Recall for *glimpse* vs *glimpseLD*.

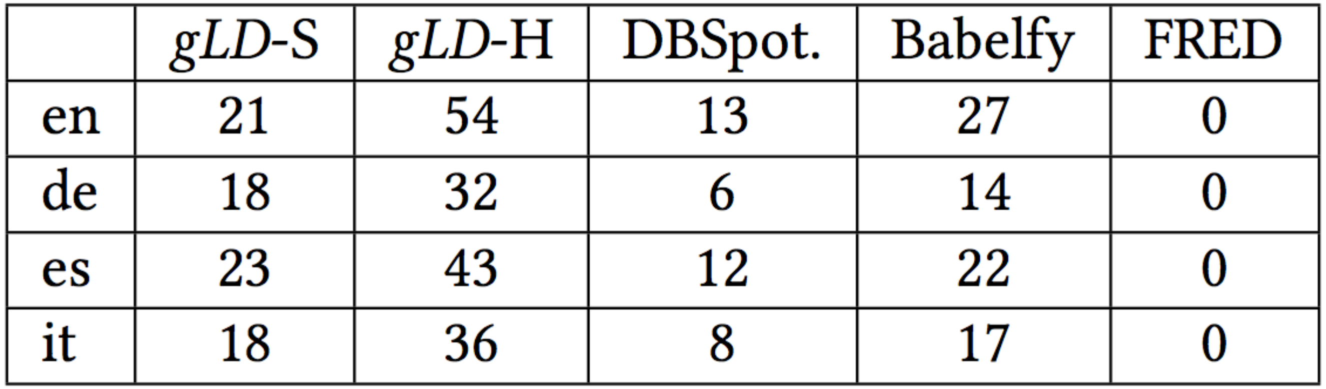
- **glimpse** → one manually provided seed
- **glimpseLD** →Linked Data seeds
- in 10 iterations **glimpseLD** can cover the same lexicon that would take more than 20 iterations with **glimpse**

Multi-lingual experiment: **Colors**

- **DATA**: Twitter stream 1st-14th of January 2016 – lang: **En**, **De**, **Es**, **It**
  - contain at least one mention of a color
    - gold standard lists of colors from Wikidata and Dbpedia
  - balance datasets size in different languages
    - 155, 828 tweets per language
- **TASK**: expand the lexicon of colors
- **user-in-the-loop:** 4 native speakers, 10 iterations

Multi-lingual experiment: **Colors**

- new color items extracted from Twitter data:
  - German: 5
  - Italian: 5
  - English: 19
  - Spanish: 22
    - azulgrana
    - rojo vivo
    - “limn" (in place of the color límon)



	gLD-S	*gLD*-H	DBSpot.	Babelfy	FRED
en	21	54	13	27	0
de	18	32	6	14	0
es	23	43	12	22	0
it	18	36	8	17	0

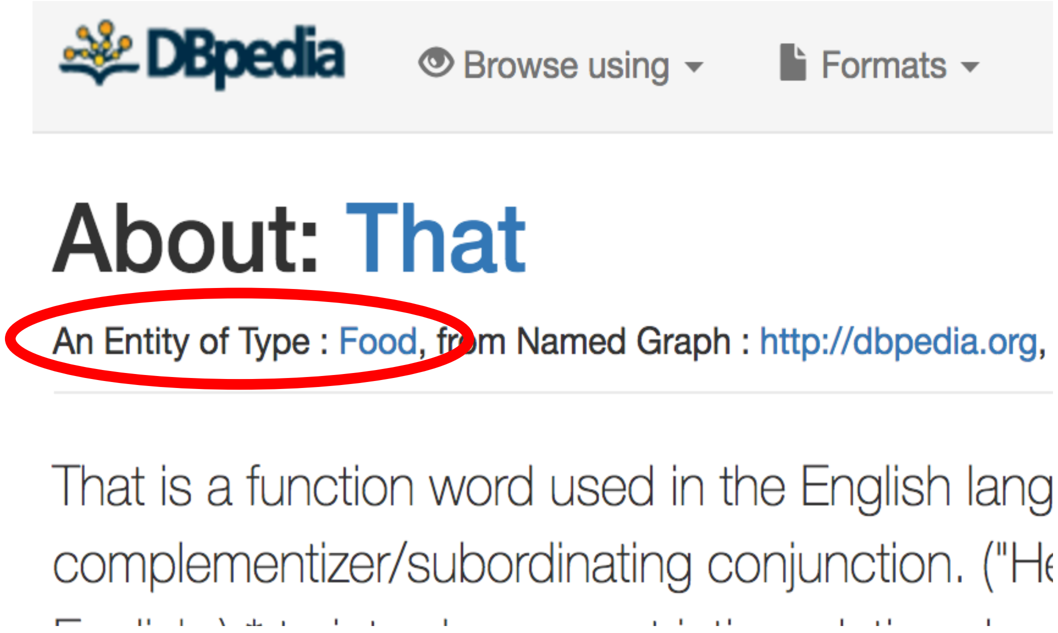
Conclusions

WHAT

- knowledge resources are never complete/exhaustive
- construct / improve dictionaries from text corpora

HOW

- iterative and purely **statistical** algorithm
  - no feature extraction required
  - **comparable** behavior for **different languages**
- organically incorporates **human feedback**

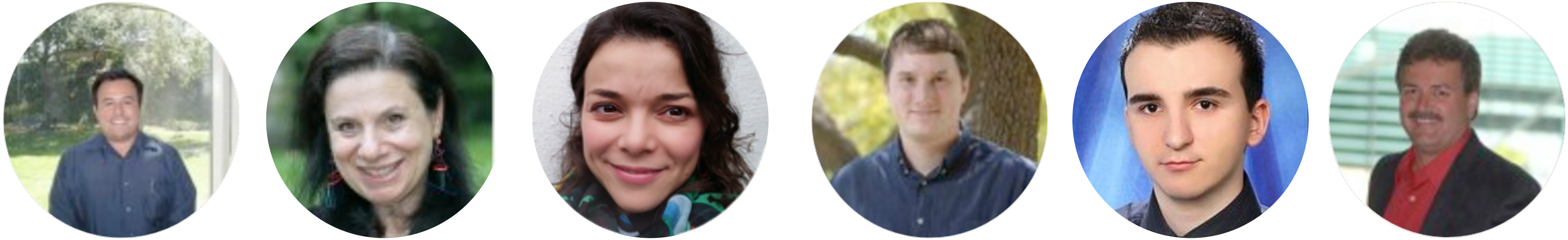


About: That

An Entity of Type : Food, from Named Graph : http://dbpedia.org,

That is a function word used in the English lang
complementizer/subordinating conjunction. ("He

Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop



Alfredo Alba, Anni Coden, **Anna Lisa Gentile**, Daniel Gruhl, Petar Ristoski, Steve Welch

annalisa.gentile@ibm.com

@AnLiGentile