How to Integrate Spark MLlib and Apache Solr to Build Real-Time Entity Type Recognition System for Better Query Understanding

Walid Shalaby, Khalifeh AlJadda,
Mohammed Korayem, Trey Grainger

About Me

Khalifeh AlJadda

Lead Data Scientist, Search Data Science

Joined CareerBuilder in 2013

PhD, Computer Science – University of Georgia

BSc, MSc, Computer Science, Jordan University of Science and Technology

Activities:

Conference Chair of Southern Data Science Conf (www.southerndatascience.com)

Founder and Chairman of CB Data Science Council

Frequent public speaker in data science and Big Data conferences.

Creator of GELATO (Glycomic Elucidation and Annotation Tool)

Search Technology at CareerBuilder by Numbers

100 million **+**
Searches per day

1,5 billion **+**
Documents indexed and searchable

500+
Search Servers

30+
Software Developers, Data Scientists + Analysts

1
Global Search Technology platform

Powering 50+ Search Experiences Including:

Search Pro
(Search - CareerBuilder RDB, Recruitment Edge, Supply & Demand)

Small Business Resume Database
(Search Basic, RDB Basic)

Talentstream Engage
(Talent Network)

Talentstream Recruit
(CareerBuilder1)

Talentstream Supply & Demand
(Supply & Demand Portal)

Candidate Sourcing Platform

Broadbean Resume Search
(Multi-vendor Resume Search)

Talentstream Gather
(Talent Gather)

CAREERBUILDER

workinretail.com

WorkInNursingJobs.com

StaffNurse.com

phonemploi.com

Lesjeudis.com

RecruLex.com

JOBSCENTRAL

erecrut.com

OILANDGAS JOB SEARCH

WorkInTherapyJobs.com

MiracleWorkers.com

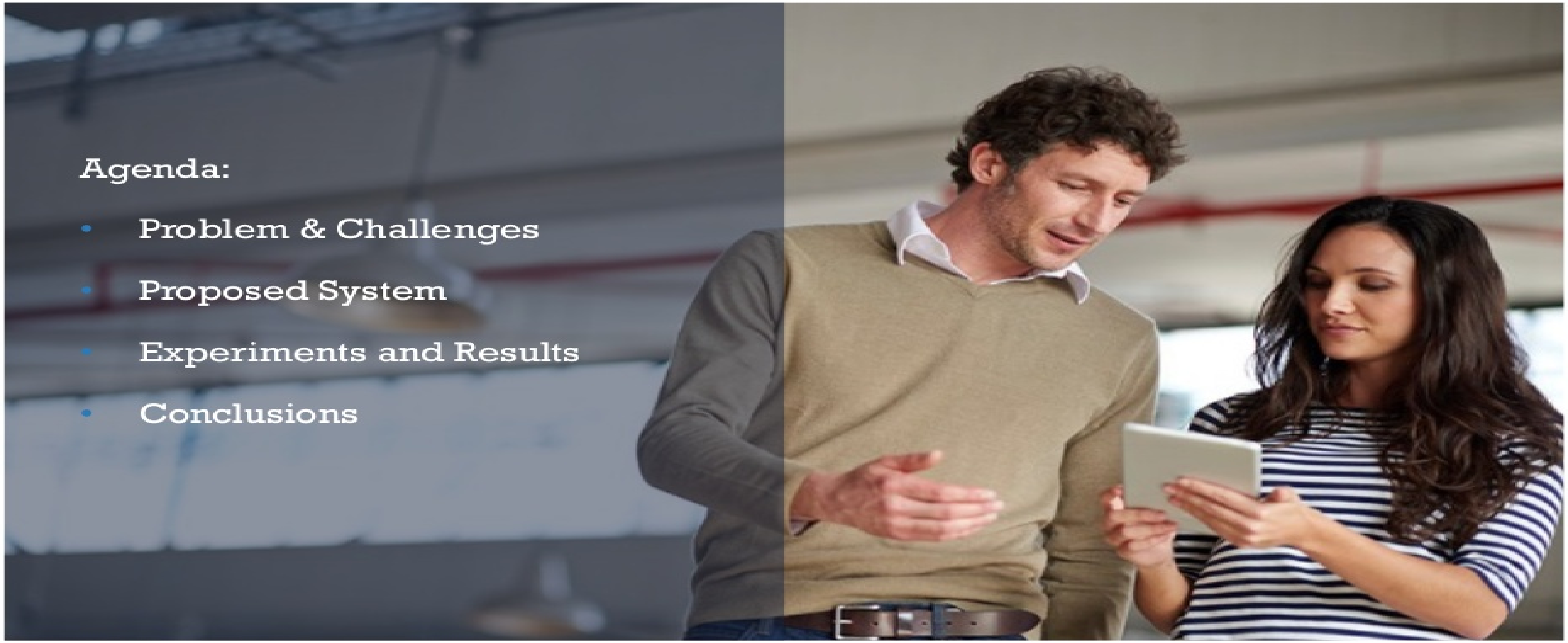
sologig.com

HEADHUNTER.com

MoneyJobs

CAO emplois.com

...and many more

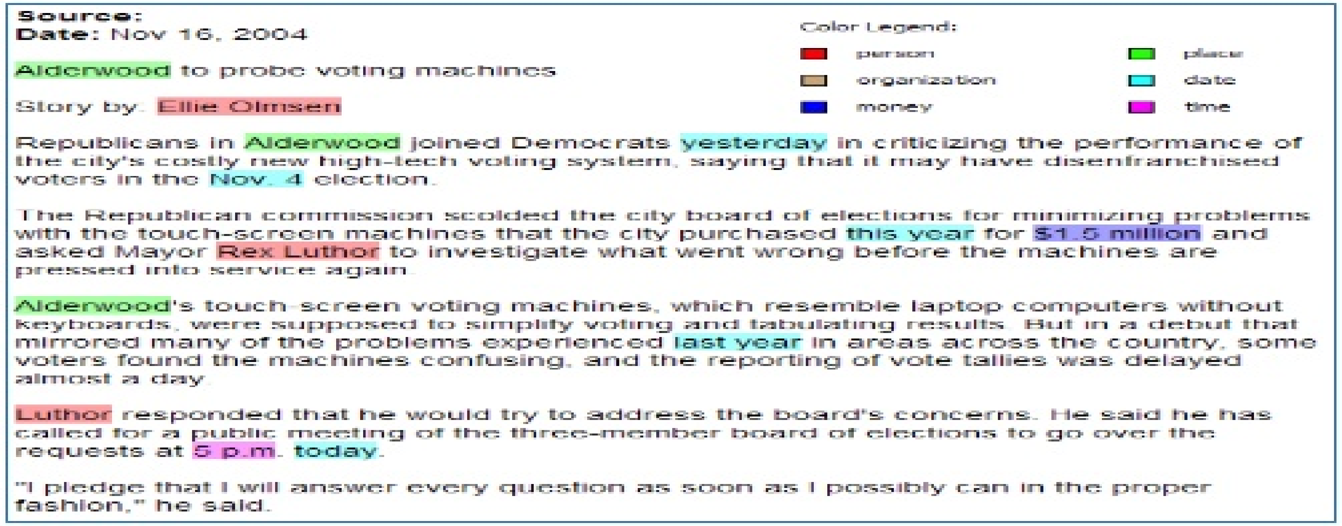


Agenda:

- Problem & Challenges
- Proposed System
- Experiments and Results
- Conclusions

Entity Type Recognition (ETR)

- Identifying regions of text corresponding to entities
- Categorizing recognized entities into predefined classes



Source:
Date: Nov 16, 2004

Color Legend:

🟥 person
🟫 organization
🟦 money
🟩 place
🟦 date
🟪 time

Alderwood to probe voting machines

Story by: Ellie Olmsen

Republicans in Alderwood joined Democrats yesterday in criticizing the performance of the city's costly new high-tech voting system, saying that it may have disenfranchised voters in the Nov. 4 election.

The Republican commission scolded the city board of elections for minimizing problems with the touch-screen machines that the city purchased this year for $1.5 million and asked Mayor Rex Luthor to investigate what went wrong before the machines are pressed into service again.

Alderwood's touch-screen voting machines, which resemble laptop computers without keyboards, were supposed to simplify voting and tabulating results. But in a debut that mirrored many of the problems experienced last year in areas across the country, some voters found the machines confusing, and the reporting of vote tallies was delayed almost a day.

Luthor responded that he would try to address the board's concerns. He said he has called for a public meeting of the three-member board of elections to go over the requests at 5 p.m. today.

"I pledge that I will answer every question as soon as I possibly can in the proper fashion," he said.

Source: Ward, Matthew O., Georges Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010.

Problem & Challenges

- Understand entity types in search queries
  - Company, Job titles, School, Skill

- Search queries are short, context-less
  - “data scientist hadoop careerbuilder”.

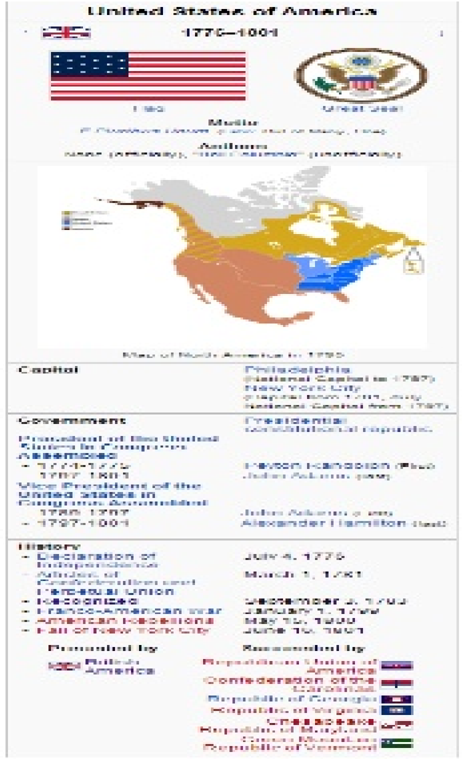
- Some entities have multiple surface forms
  - university of north carolina charlotte, unc charlotte, uncc

- Domain specific entities
  - registered nurse (rn), licensed practical nurse (lpn), director of nurse (don)

Prior Work

- Wikipedia, Wikipedia, Wikipedia
  - Title and first paragraph
  - Title and categories
  - Infobox

- DBpedia, Freebase



United States of America

1776–1801

Flag | Great Seal

Motto

[illegible]

Anthem

[illegible]

Map of North America in 1795

Capital	Philadelphia (National Capital to 1787) New York City (Capital from 1785, only National Capital from 1787)
Government	Presidential constitutional republic
President of the United States in Congress Assembled	
• 1776–1779	Peyton Randolph (first)
• 1797–1801	John Adams (last)
Vice President of the United States in Congress Assembled	
• 1785–1787	John Adams (first)
• 1797–1801	Alexander Hamilton (last)
History	
• Declaration of Independence	July 4, 1776
• Articles of Confederation and Perpetual Union	March 1, 1781
• Recognized	September 3, 1783
• Franco-American War	January 1, 1799
• American Rebellions	May 15, 1800
• Fall of New York City	June 15, 1801
Preceded by	**Succeeded by**
British America	Republican States of America Confederation of the Carolinas Republic of Georgia Republic of Virginia Chesapeake Republic of Maryland Green Mountain Republic of Vermont

Limitations

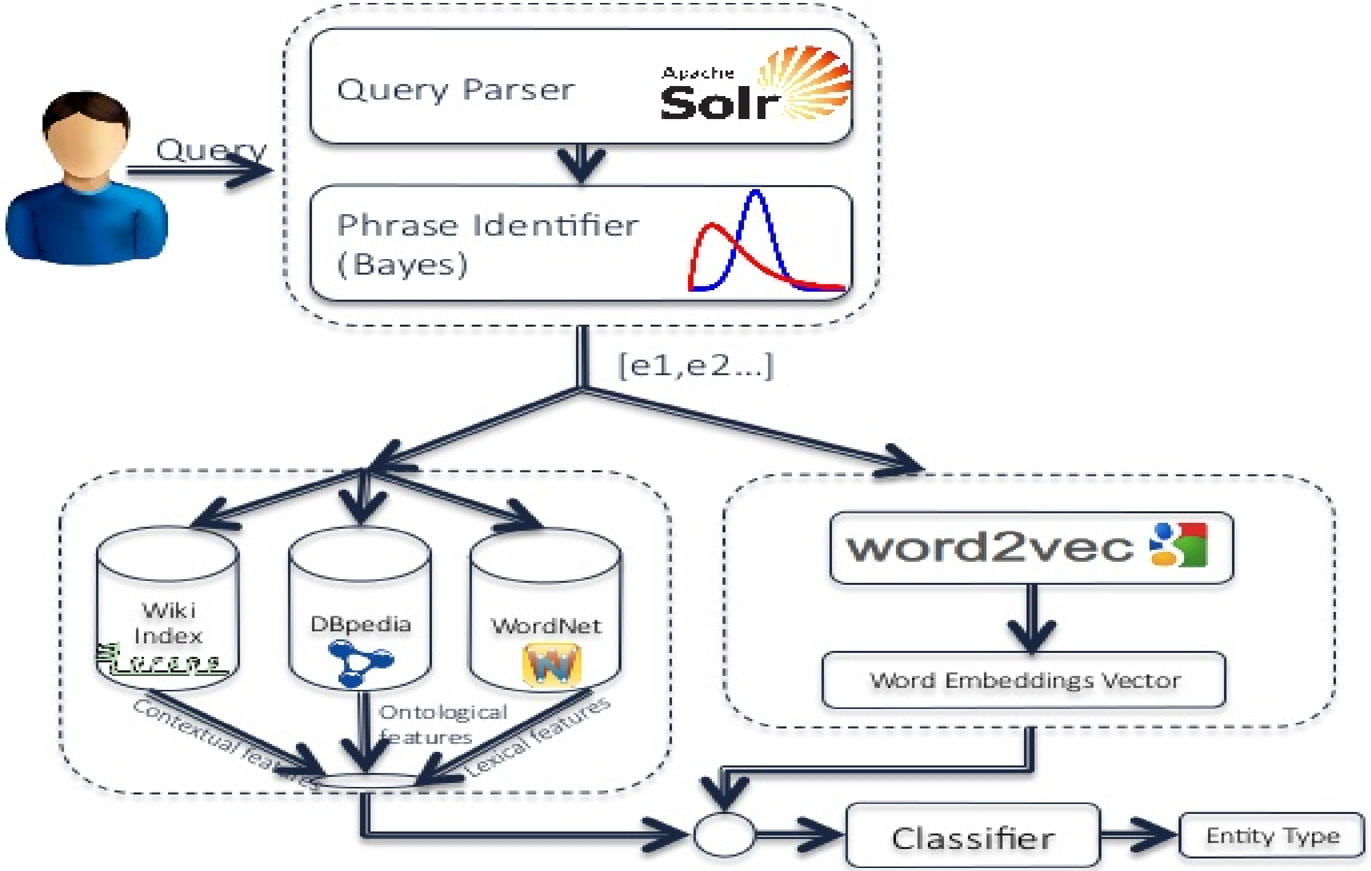
- Entities with no page/entry (java developer)
- Non-standard categories (skill)
- Domain specific knowledge (e.g., job posts)

Methodology

Enrich entity representation using 4 clues (features):

- Real-world knowledge (Wikipedia)
- Domain specific knowledge (job posts)
- Ontology (DBpedia)
- Lexical DB (Wordnet)

Architecture



CAREERBUILDER

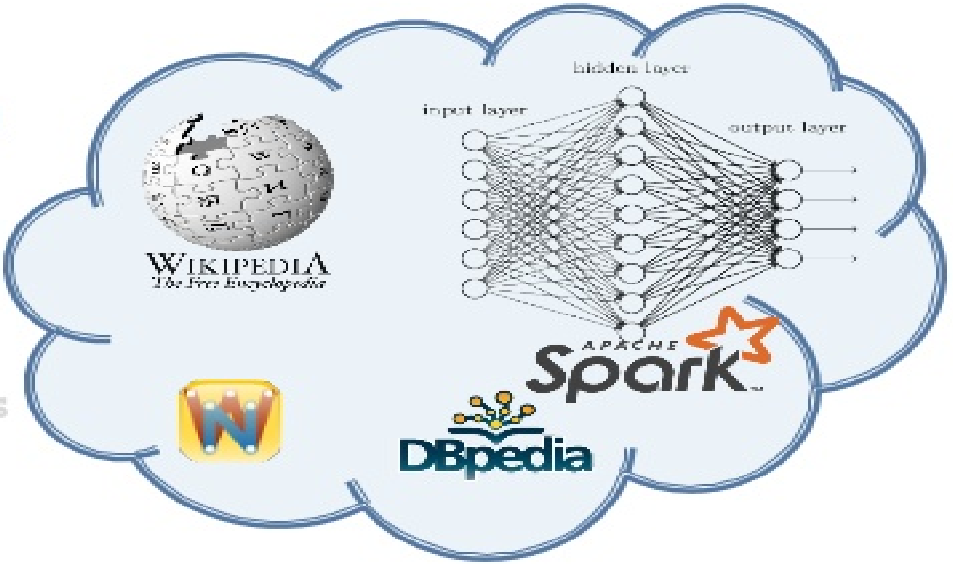
Our System

Offline

- Wikipedia index (title, length, text, categories)
- Word2Vec trained on ~100M job posts
- SVM classifier

Online

- top 3 hits + 10 fragments for each + categories
- Word2Vec synset (most similar 20 words)
- tf-idf vector
- is_company (binary feature)
- is_agent_noun (binary feature)



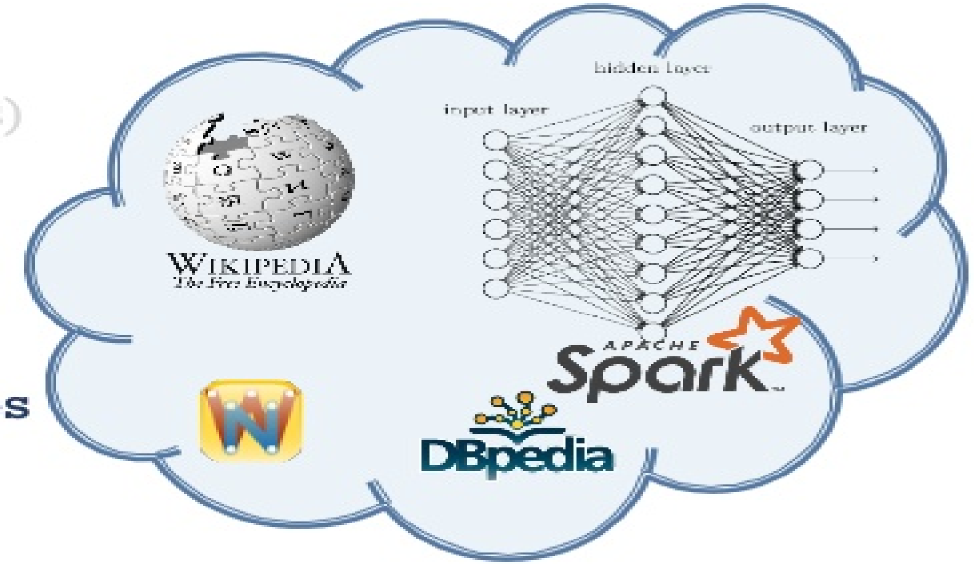
Our System

Offline

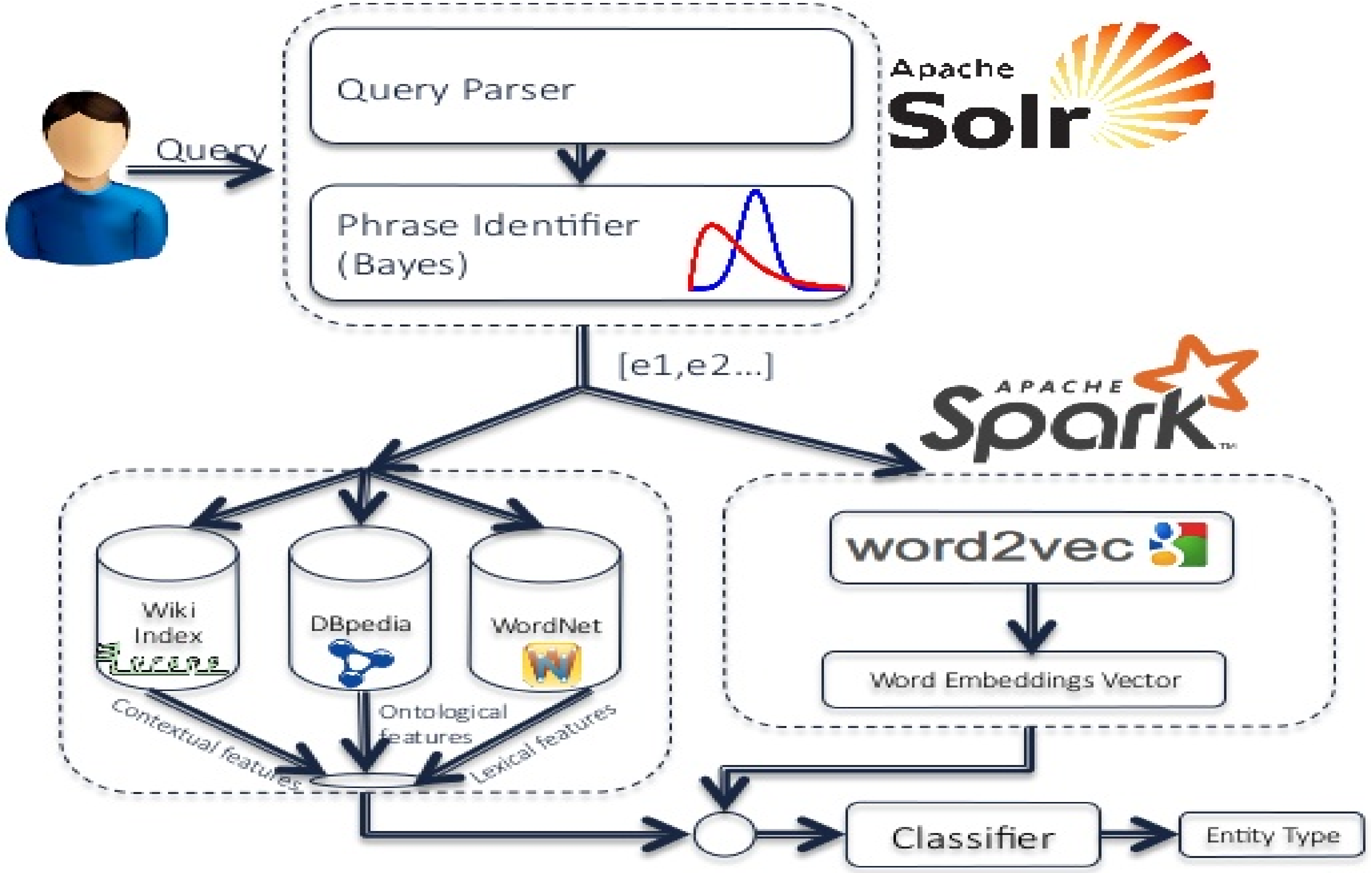
- Wikipedia index (title, length, text, categories)
- Word2Vec trained on ~25m job posts
- SVM classifier

Online

- top 3 hits + 10 fragments for each + categories
- Word2Vec synset (most similar 20 words)
- tf-idf vector
- is_company (binary feature)
- has_agent_noun (binary feature)

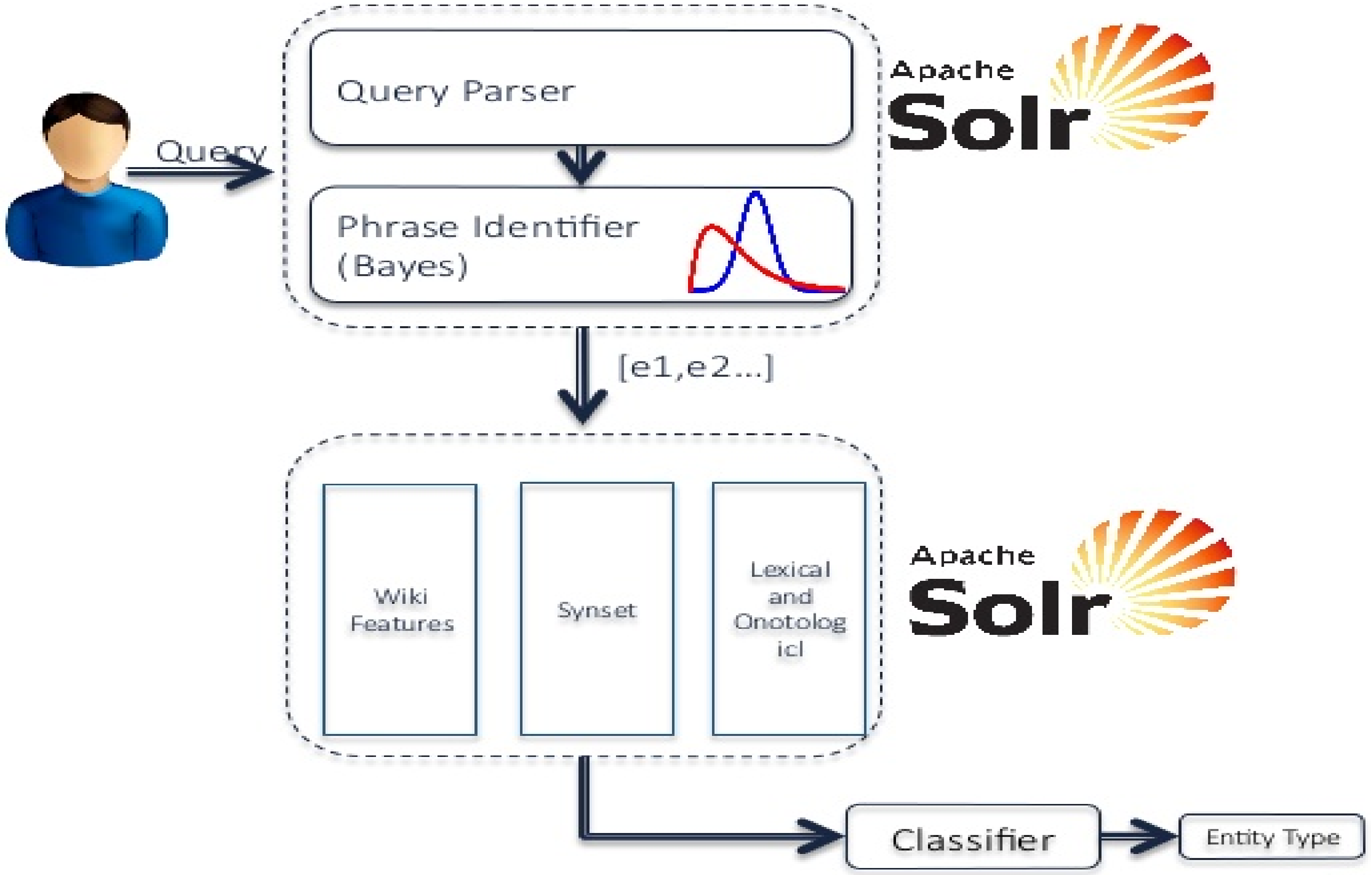


Offline Architecture



CAREERBUILDER

Online Architecture



Context Features

- **IBM <company>**

The International Business Machines Corporation (**IBM**) is an American **multinational** technology and **consulting corporation**, with **headquarters** in Armonk, New York. **IBM manufactures** and **markets** computer hardware had originated with CTR's Canadian **subsidiary**. The initialism **IBM** followed. Securities analysts. In 2012, 'Fortune' ranked **IBM** the No. 2 largest U.S. **firm** in terms of number of employees (435,000 ('Barron's'), No. 5 most admired **company** ('Fortune'), and No. 18 most innovative **company** ('Fast **Company**'). The **company** held the record to Lenovo (2005, 2014). In 2014 **IBM** announced that it would go 'fabless' by offloading **IBM** Micro Electronics form the core of what would become International Business Machines (**IBM**). Julius E. Pitrat patented was renamed the 'International Business Machines **Corporation**' (IBM), citing the need to align its name the **company** produced small arms ...

- **LPN <job title>**

Lee Presson and the Nails (also known as **LPN**) is a **swing band** that formed in the San Francisco. As of 2010, the **band** has released five **albums**. **LPN** differentiated themselves from of band leader **Lee Presson...**

Embedding Features (Synsets)

- **IBM <company>**

fusion iis jms virtualization
emc weblogic atg esb
mq nosql voip tibco jboss
Tivoli hadoop avaya citrix
tomcat hp websphere

- **LPN <job title>**

practitioner lvn nurse
registered rn vocational
psychologist midwife aide can
licensed licensure icu
psych arnp

Evaluation

- 2 baselines vs. combinations of various representations
- 10 fold cross-validation
- Report P, R, and micro-averaged F1
- Data set (177K entities)

Category	Number of instances
Company	40000+
Job Title	3500+
School	100000+
Skill	25000+

Baseline Results

Category/Model	Company	Job Title	School	Skill
Bow	85.19	87.42	96.59	76.57
$wiki_w$	5.35	2.24	1.06	11.18

- Absolute increase (F1)
- Absolute decrease (F1)

Context Features Results

Category/Model	Company	Job Title	School	Skill
Bow	85.19	87.42	96.59	76.57
$wiki_w$	5.35	2.24	1.06	11.18
$wiki_x$	10.79	-0.14	1.93	15.64

- Absolute increase (F1)
- Absolute decrease (F1)

Context + Embedding Features Results

Category/Model	Company	Job Title	School	Skill
Bow	85.19	87.42	96.59	76.57
$wiki_w$	5.35	2.24	1.06	11.18
$wiki_x$	10.79	-0.14	1.93	15.64
$wiki_x + job_w$	10.99	2.70	2.09	16.24

- Absolute increase (F1)
- Absolute decrease (F1)

Context + Embedding + Ontology + Lexical Features Results

Category/Model	Company	Job Title	School	Skill
Bow	85.19	87.42	96.59	76.57
$wiki_w$	5.35	2.24	1.06	11.18
$wiki_x$	10.79	-0.14	1.93	15.64
$wiki_x + job_w$	10.99	2.70	2.09	16.24
$wiki_x + job_w + ont + lex$	11.37	3.15	2.10	16.57

- Absolute increase (F1)
- Absolute decrease (F1)

Conclusion

- Effective approach for ETR of search query entities
- Tailored to the job search and recruitment domain
- Combine real-world and domain-specific knowledge
- The ensemble entity representation
  - contextual information using Wikipedia
  - semantic information in millions of job postings
  - class type in DBpedia for Company entities
  - linguistic properties in WordNet for Job Title entities
- Ensemble features gave 97% micro-averaged F1
- Online ETR takes 30ms per entity type request

Thank you!

www.aljadda.com
Twitter: @aljadda

CAREER BUILDER™