Synthetic Data Generation for Natural Language Understanding with Probabilistic Context Free Grammars

Georgios Balikas, Lead Data Scientist
Twitter: @gbalikas

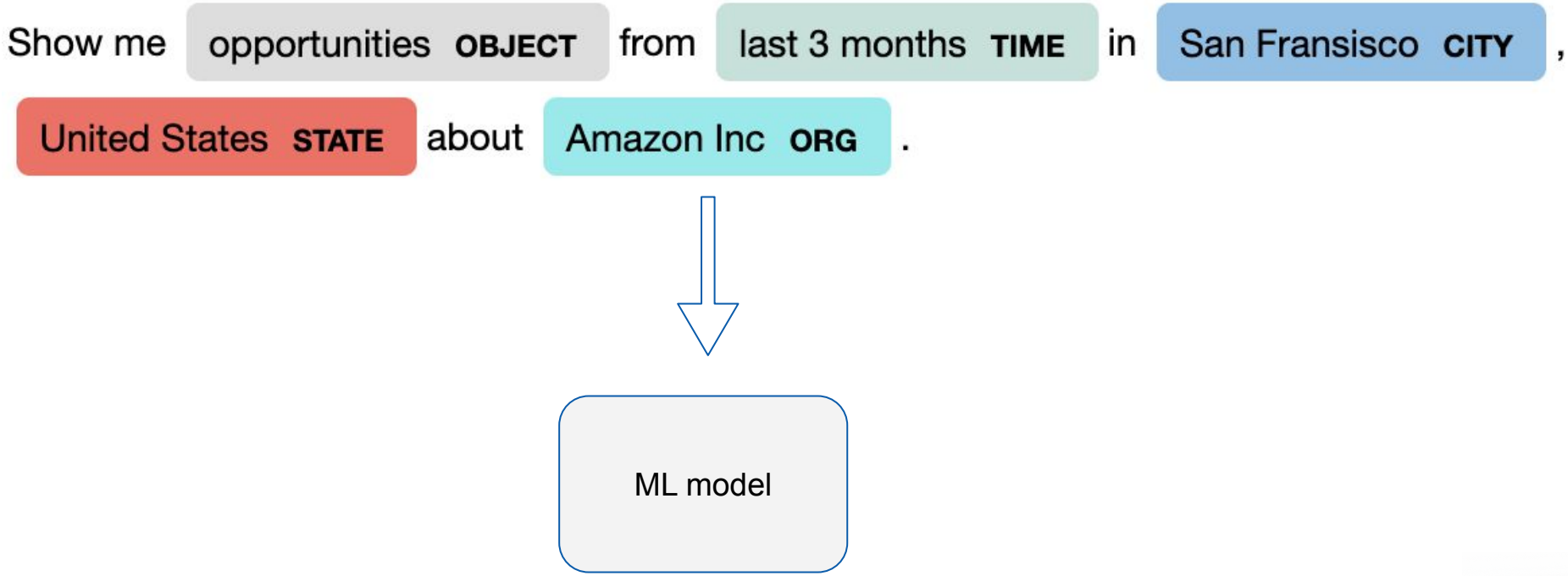
Forward-Looking Statements

This presentation contains forward-looking statements about, among other things, trend analyses and future events, future financial performance, anticipated growth, industry prospects, environmental, social and governance goals, and the anticipated benefits of acquired companies. The achievement or success of the matters covered by such forward-looking statements involves risks, uncertainties and assumptions. If any such risks or uncertainties materialize or if any of the assumptions prove incorrect, Salesforce's results could differ materially from the results expressed or implied by these forward-looking statements. The risks and uncertainties referred to above include those factors discussed in Salesforce's reports filed from time to time with the Securities and Exchange Commission, including, but not limited to: the impact of, and actions we may take in response to, the COVID-19 pandemic, related public health measures and resulting economic downturn and market volatility; our ability to maintain security levels and service performance meeting the expectations of our customers, and the resources and costs required to avoid unanticipated downtime and prevent, detect and remediate performance degradation and security breaches; the expenses associated with our data centers and third-party infrastructure providers; our ability to secure additional data center capacity; our reliance on third-party hardware, software and platform providers; the effect of evolving domestic and foreign government regulations, including those related to the provision of services on the Internet, those related to accessing the Internet, and those addressing data privacy, cross-border data transfers and import and export controls; current and potential litigation involving us or our industry, including litigation involving acquired entities such as Tableau Software, Inc. and Slack Technologies, Inc., and the resolution or settlement thereof; regulatory developments and regulatory investigations involving us or affecting our industry; our ability to successfully introduce new services and product features, including any efforts to expand our services; the success of our strategy of acquiring or making investments in complementary businesses, joint ventures, services, technologies and intellectual property rights; our ability to complete, on a timely basis or at all, announced transactions; our ability to realize the benefits from acquisitions, strategic partnerships, joint ventures and investments, including our July 2021 acquisition of Slack Technologies, Inc., and successfully integrate acquired businesses and technologies; our ability to compete in the markets in which we participate; the success of our business strategy and our plan to build our business, including our strategy to be a leading provider of enterprise cloud computing applications and platforms; our ability to execute our business plans; our ability to continue to grow unearned revenue and remaining performance obligation; the pace of change and innovation in enterprise cloud computing services; the seasonal nature of our sales cycles; our ability to limit customer attrition and costs related to those efforts; the success of our international expansion strategy; the demands on our personnel and infrastructure resulting from significant growth in our customer base and operations, including as a result of acquisitions; our ability to preserve our workplace culture, including as a result of our decisions regarding our current and future office environments or work-from-home policies; our dependency on the development and maintenance of the infrastructure of the Internet; our real estate and office facilities strategy and related costs and uncertainties; fluctuations in, and our ability to predict, our operating results and cash flows; the variability in our results arising from the accounting for term license revenue products; the performance and fair value of our investments in complementary businesses through our strategic investment portfolio; the impact of future gains or losses from our strategic investment portfolio, including gains or losses from overall market conditions that may affect the publicly traded companies within our strategic investment portfolio; our ability to protect our intellectual property rights; our ability to develop our brands; the impact of foreign currency exchange rate and interest rate fluctuations on our results; the valuation of our deferred tax assets and the release of related valuation allowances; the potential availability of additional tax assets in the future; the impact of new accounting pronouncements and tax laws; uncertainties affecting our ability to estimate our tax rate; uncertainties regarding our tax obligations in connection with potential jurisdictional transfers of intellectual property, including the tax rate, the timing of the transfer and the value of such transferred intellectual property; uncertainties regarding the effect of general economic and market conditions; the impact of geopolitical events; uncertainties regarding the impact of expensing stock options and other equity awards; the sufficiency of our capital resources; our ability to comply with our debt covenants and lease obligations; and the impact of climate change, natural disasters and actual or threatened public health emergencies, including the ongoing COVID-19 pandemic.

Outline

- Named Entity Recognition
- PCFG
- Synthetic data with PCFG

NER: locate and classify text spans in unstructured text into pre-defined categories

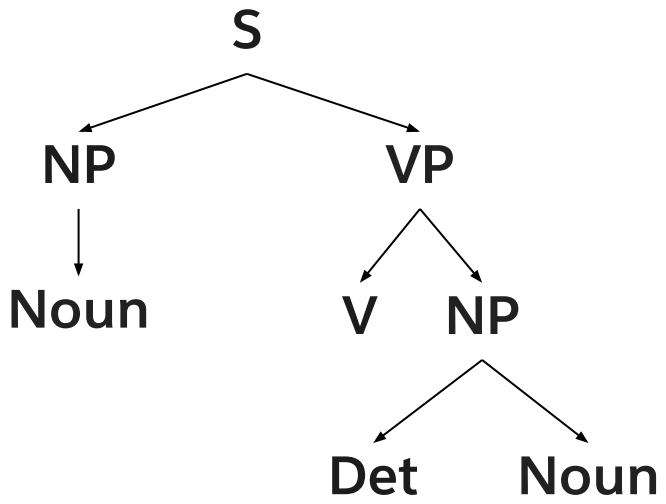


We need to be able to generate **custom NER training data** for the **domain specific models** that power our applications

Probabilistic Context Free Grammars allow us to describe sequences of contexts

A simple PCFG
1.0 S ->NP VP
0.2 NP -> Adj Noun
0.7 NP -> Det Noun
0.1 NP -> Noun
1.0 VP -> V NP
.
.

Example

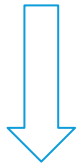


Potatoes [Noun] cooked [V] in [Det] oven [Noun]

Probabilistic Context Free Grammars allow us to describe sequences of contexts

A simple PCFG
1.0 S -> Pre Obj Suf
0.2 Pre -> ACTION PER
0.7 Obj -> OPP | ACC
Suf -> LOC TIME
LOC -> CITY STATE COUNTRY
TIME -> ...
.
.
.

action.vocab
per.vocab
org.vocab
city.vocab
...



Show me [Action] opportunities [Obj] from last 3 months [TIME] in [PREP] San Fransisco [CITY], United States [Country] about Amazon Inc [ORG].

Concluding remarks

- For NLU there is a need to create custom data respecting some properties.
- Frequently, there is some domain knowledge on how these data should look like e.g., sequence of tags
- There is some notion of probability of occurrence too for words that comprise each tag
- PCFGs can be used to create such data massively using sampling mechanisms
- We have successfully used this mechanism to create training data for semantic parsing on a system that is now in production at Salesforce.

More: “Query Understanding for Natural Language Enterprise Search” DeepNLP@SIGIR‘20


Thank You

3D Astro

GET MORE 3D ASTROS

3D Astro

GET MORE 3D ASTROS

3D Einstein

GET MORE 3D EINSTEINS

3D Cloudy and Misc.

GET MORE 3D CLOUDYS