Apposition extraction, a difficult problem
James Curran and Will Radford

A study by CMCRC researchers presents a fresh look at extracting apposition from large collections of news, web and broadcast text in order to turn unstructured news stories into "computable data". News is about interactions between **entities** - people, places and organisations - and understanding stories requires interpreting the entities in them and their attributes.

Apposition is a linguistic construction often used by writers to introduce entity attributes and many automated systems have taken advantage of it to extract information about entities. Researchers Will Radford and Professor James Curran find that, despite the common perception that analysis of commas and the grammatical structure is sufficient for high accuracy, apposition extraction is a difficult challenge.

The sentence below mentions an entity, John Ake, his age and role -- all information that paint a clearer picture of him. These types of information are diverse and the role portion itself contains apposition, stating that "American Capital Management & Research Inc." is based in "Houston".

> **John Ake**, *48*, *a former vice-president*
> *in charge of legal compliance at American Capital*
> *Management & Research Inc., in Houston*, ...

While identifying apposition is straightforward for a human reader, it can be surprisingly hard for machines. Correctly identifying attribute spans can be difficult - which comma should mark the right-hand boundary? As a person can be a "vice-president", a company cannot, but they can both be a "contender". Judging where an attribute is compatible with the entity requires an understanding of world knowledge, a challenge for any automated system.

Researchers Will Radford and Professor James Curran develop three techniques based on better grammatical, semantic and statistical modelling to automatically identify apposition in text. The general framework identifies candidate phrases, entities and their attributes, then classifies them as apposition using machine learning. Rules based on the linguistic theory of apposition filter the candidates by their grammatical structure. Semantic features encode the fact that "vice-president" is a valid role for a person entity. Finally, the statistical model describes the pair of entity and information together where previous approaches consider them separately. These complementary approaches combine for a more faithful modelling of apposition and result in a system with state-of-the-art performance.

Apposition extraction is just one feature of language that the CMCRC's text analytics team is exploiting to unlock data from text. These data can then be used for a variety of applications: linking ambiguous entities in online news for algorithmic trading; extracting board member information from news stories for electronic corporate governance, or identifying alternative medical terminology for health insurance claim processing. Better language processing is critical wherever you need to understand entities.