I Read The News Today, Oh Boy

Making Metaphors Topical, Timely and Humorously Personal

Tony Veale, Hanyang Chen, Guofu Li

School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin D4, Ireland
Tony.Veale@UCD.ie, Hanyang.Chen@ucdconnect.ie

Abstract. Human speakers do not create metaphors in a vacuum. Our rhetorical urges are tempered by a variety of contextual factors, such as *ethos* (does a metaphor reflect my values?), *relevance* (does a metaphor speak to my topic?), *timeliness* (is this a good time to use this metaphor?) and *affect* (does this metaphor stir the desired emotions in my audience?). The 24-hour news cycle offers an ideal setting in which to explore automated metaphor generation that is both timely and topical, as not only do journalists rely on pithy metaphors to attract readers, readers often respond to the news with wittily apt, conversation-sparking metaphors of their own. Indeed, as micro-blogging platforms such as Twitter provide digital printing presses for the masses that also allow us to turn our lives and opinions into 140-character headlines, we can use computational techniques to craft personalized metaphors that suit a specific human recipient. In this paper we explore metaphor generation techniques that are shaped for a specific topical context, using approaches to topic modeling such as Latent Dirichlet Allocation, or that reflect the online personality of a specific recipient, as evidenced by their most recent emations or *tweets*. Each approach is instantiated in an autonomous *Twitterbot*, a system that creates and tweets its own content without human curation. We use Twitterbots to study the potential for humour to arise from the timely online interaction of humans and machines.

Keywords: Metaphor, Topicality, Affect, Personality, News, Humour, Twitter

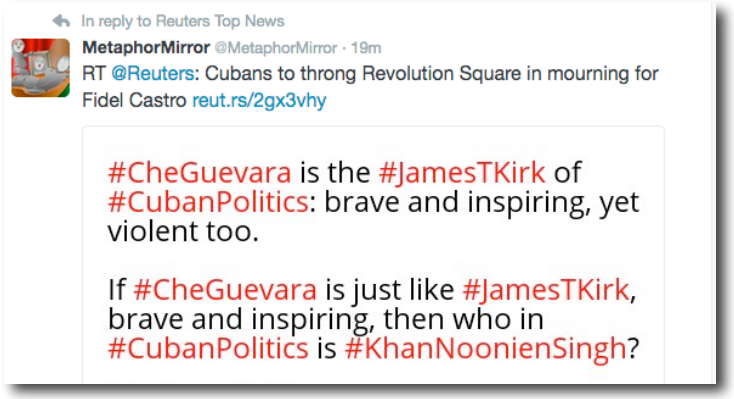
1 Metaphor Mirror on the wall

We want the news to hold up a mirror to world events, yet we are not so naïve as to expect this mirror to be without bias or distortion. For the news does more than *report*: it shapes our view of events, by telling us where to look, what to see and oftimes what to *think*. So to readers at one end of the political spectrum, the news emanating from the opposing end can resemble the reflection of a funhouse mirror, leading readers to seek out those providers whose biases and distortions accord with their own. For a balanced view of world events, we can obtain our news from diverse sources, yet jumping between providers of very different orientation or register on the Web can be a jarring source of cognitive dissonance. Nonetheless, to have one's tacit expectations of the news and of newsmakers laid bare in this way is also a source of insight that occasionally rises to the level of what Koestler (1964) calls a *bisociation*.

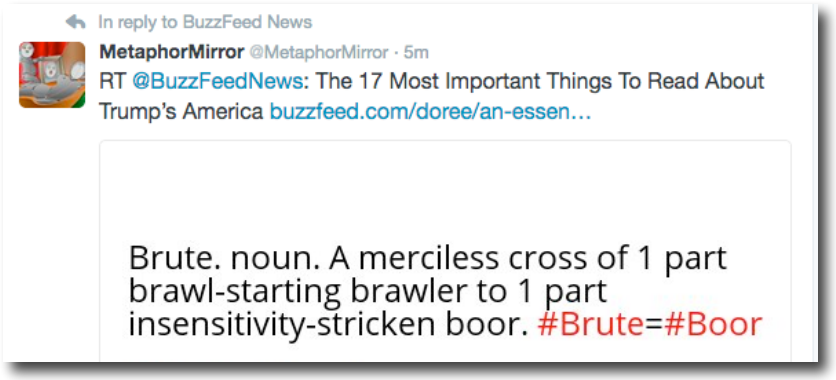
Reporters use metaphor to concisely frame current events from a certain affective

perspective. Thus, to say that "*Google's halo has been tarnished*" by recent events is to suggest that many – perhaps the company itself – see Google as a "saint", while to say that "*Oracle's crown has slipped*" is to portray the company as the key player in its field. Such metaphors achieve a subtle coercion, insomuch as they presuppose much more than they are willing to put into words. To *break set* and see beyond any particular framing, we must challenge our received metaphors to invent counter-metaphors (and script oppositions) of our own. This is what it means to engage with the news: not to uncritically treat headlines as bundles of propositions to be added to our individual stores of knowledge, but to imagine how events might look if framed from an opposing perspective. When we debate the news with colleagues and friends, we rely as much on metaphors as on facts to examine our feelings and reach our own conclusions. Lakoff and Johnson (1980) have persuasively championed a conceptual view of metaphor that sees most linguistic metaphors as surface elaborations of deep conceptual schemas such as *Argument is War*, *Life is a Journey*, *Politics is a Game* and *Theories are Buildings*, and it is natural for readers to respond with elaborations of these schemas whenever they underpin the meaning of a headline. Yet the most challenging metaphors are those that use a different conceptual schema, to show that there are multiple sides to the same story, more than a single headline or news source can show. In this paper we set ourselves the task of creating metaphors for incoming headlines on Twitter, so that the news offerings of *@nytimesworld, @FOXnews* and *@CNNbrk* can be automatically paired with original metaphors that prompt readers to imagine different, perhaps humorously different, viewpoints on the same story.

Like jokes, metaphors thrive on semantic tension, both in themselves (between *source* and *target* domains) and between the metaphors themselves and their contexts of use. A challenging metaphor should be apt yet surprising, and exhibit what Attardo (2001) calls *relevant inappropriateness* or what Oring (2003) calls *appropriate incongruity*. Ideally, the pairing of metaphor to headline should create what Koestler (1964) calls a *bisociation*, a jarring but meaningful clash of overlapping frames of reference. Pollio (1996) suggests that while persuasive metaphors successfully hide the rift between source and target domains, jokes draw our attention to this rift and revel in its potential to swallow rational thought. We aim for the machine-generated metaphors selected by *@MetaphorMirror*, a Twitterbot that pairs novel metaphors to incoming news headlines, to achieve both of these ends: to appear appropriate to their contexts of use while also hinting at the conflicts of ideas and world views that lurk behind the news. Consider the following pairing of metaphor to headline that was tweeted by our Twitterbot *@MetaphorMirror* on the day Fidel Castro died:



The stronger the language of the metaphor, the more humorous the perceived opposition with its target headline. Consider this metaphorical take on the news:



Metaphors such as these can be generated using a hybrid knowledge-driven and data-driven approach that relies on a mix of symbolic and statistical knowledge. In the case above, symbolic knowledge is used to establish and then pithily articulate the overlap between the concepts *Brute*, *Brawler* and *Boor*, while statistical knowledge is needed to map this textual formulation to a headline about the president-elect Donald Trump. Specifically, a robust statistical approach to topic modelling is required to see the non-literal similarities between Mr. Trump specifically – a man whose aggressive and boorish manners are the subject of many newspaper articles, among them those that actually support his agenda – and the stereotypical idea of a brute in a bar brawl. Our goal in this paper is to describe the workings of this hybrid model and to show how the pairing of the topically literal and the metaphorically apt can give rise to script oppositions that are both timely and humorous.

With these goals in mind, the rest of the paper assumes the following structure: section 2 considers the automated generation of metaphors, showing how the creation process must combine rote knowledge with inspiration; section 3 explores the news domain, to model the news as its own conceptual space that can be mapped, via a characterization of topics, onto the space of machine-generated metaphors; section 4 then describes *@MetaphorMirror*, a creative bot that applies this mapping to Twitter news; section 5 presents an empirical evaluation of *@MetaphorMirror*'s pairings of metaphors to headlines; and finally, section 6 shows this approach to topical metaphor generation can be extended to look beyond the content of the major news providers to the emantions of prolific individuals, such as *@realDonaldTrump*. From there we show that our metaphors need not always be keyed to the topics of a news feed, and can instead be formulated to suit the online personalities of individual Twitter users.

2 Metaphors: Shaped By Knowledge and Inspired By Data

The cleverest metaphors are for naught if they cannot be understood by their intended audience, and so it makes good communicative sense to view metaphor generation and interpretation as flip sides of the same process of creative meaning-making. It

makes just as much computational sense for us to model automated generation as the flip side of automated interpretation, and to simply apply existing theories of metaphor interpretation *in reverse* so as to allow machines to create novel metaphors of their own. A recent survey by Veale *et al.* (2016) divides computational theories into four interpretation-oriented groupings: the *corrective*, typified by Wilks (1978) and Fass (1991), see metaphor as a semantic anomaly from which an interpretation system must recover a non-anomalous meaning; the *categorial*, typified by Way (1991) and Glucksburg (1998), see metaphor as enlarging one's category system by finding a new place in a new category for an idea exhibiting key features of that category; the *analogical*, typified by Gentner *et al.* (1989) and Veale & O'Donoghue (2000), who posit that the chasis of a sound metaphor is a structure-mapping analogy between two domains; and the *schematic*, as typified by Lakoff and Johnson (1980), Carbonell (1981) and Hobbs (1981), who see metaphors as surface manifestations of deeper conceptual metaphors which are, in turn, anchored in embodied conceptual schemas for the mind, for emotions, for purposeful action and so on.

Each of these types of approach, though chiefly focused on metaphor interpretation, can in principle be applied to the problem of metaphor generation. For instance, Shutova (2010) presents a statistical approach to metaphor interpretation by paraphrasing, wherein an unconventional metaphorical form is rewritten in more conventional language (e.g. "she *swallowed* her anger" becomes "she *suppressed* her anger"). Though largely corrective, many of the same statistical mechanisms can be applied in reverse to paraphrase normative language using *less* conventional phrasing (e.g., see also Harmon, 2015). Veale and Li (2013) present a means of building the kind of dynamic, fine-grained category hierarchy presupposed in the theories of Way and Glucksberg, using information extraction from the web to achieve the necessary scale and diversity. They also demonstrate the utility of their web service, named *Thesaurus Rex*, for understanding *and* generating metaphors via category membership norms. Veale & Li (2011) further show how propositions can be extracted from the *why do* questions that are found in the query logs of popular search engines (or in the query completions offered by these engines), and demonstrate how analogical mapping can be performed over this structured content. Though the most potent schematic structures of Lakoff & Johnson's (1980) conceptual theory have been inventoried in the *Master Metaphor List* (Lakoff, 1994), it is also possible to extract commonplace schemas using automated corpus analysis (e.g. see Mason, 2004 and, to a lesser extent, Harmon, 2015). Veale (2015) uses the Google n-grams (Brants & Franz, 2006) to find *is-a* statements with the potential to serve as schemas (such as the 4-gram "*crime* is a *disease*") and uses a mix of property-level knowledge (from Veale & Hao, 2007), propositional content (from Veale & Li, 2011) and taxonomic structures (from Veale & Li, 2013) to filter and instantiate these *pseudo*-schemas in novel but meaningful metaphors.

Reiter & Dale (2006) argue that the generation of complex natural-language artifacts requires two levels of planning: macro-planning (what is it I want to say?) and micro-planning (how do I go about saying it?). While cleaving to this dichotomy, Veale (2015) uses three levels of planning: the *macro*-level (what is the main conceit of the metaphor?); the *macro-micro* (how is this conceit elaborated in a propositional form?); and the micro-level (how is this propositional form to be rendered in English?). As the metaphors in question are to be tweeted by an automated Twitterbot

without human curation, the *macro-micro* and *micro* levels explicitly concern themselves with a search for propositions and for linguistic forms that will ultimately yield a pithy remark that can be rendered in 140 characters of less. Veale (2015) describes this search as a language game (a "game of tropes") and describes a Twitterbot named *@MetaphorMagnet* that employs a wide range of tropes and rhetorical strategies (on the order of 40) to achieve a diversity of outputs that the bot's followers will find interesting and *re*-tweetable.

As *@MetaphorMagnet* uses the Google n-grams to guide its macro-level planning, and uses its various databases of stereotypical properties and behaviours to filter and elaborate its macro plans into workable semantic forms, its designers claim its actions are inspired by data but shaped by knowledge. Consider the following output from *@MetaphorMagnet* (all of whose outputs are visible on Twitter):

> When it comes to the masterpieces they produce,
> some masters can be far from beloved and can be downright lonely.
>
> Lonely masters produce eerie masterpieces the way
> wolves produce howls. #Master=#Wolf #Masterpiece=#Howl

This double-tweet metaphor is sparked at the macro-level by the Google 2-gram "*eerie masterpieces*", which prompts the planner to consider juxtaposing *masterpiece* with a known stereotype of *eeriness*. Finding *howl* in its database of stereotypical norms, the system searches the propositional contexts of *masterpiece* and *howl* to find an analogy that can cleanly map one context onto another. A concern for efficiency leads it to use predicate identicality as a matching criterion (Gentner *et al*., 1989), and so the propositions *produce*(masters, masterpieces) and *produce*(wolves, howls) are found to meet this basic requirement for a well-formed analogy. But the analogy is favored for another important reason: as the system believes *masters* are typically *beloved* and *wolves* are typically *lonely*, the antonymy of *lonely* and *beloved* adds a savory dash of semantic tension to the mix. The resulting analogy, which is in part derived from a stereotypical *dis*-analogy, is then tweeted in two parts using the framing device above. The metaphor is inspired by the contingent observation that some masterpieces are *eerie* – this fact is not found in the system's own knowledge – and semantically well-formed due to the careful planning of the *macro-micro* level, yet any pragmatic resonances in the minds of readers are mostly unplanned. Precisely why a masterpiece is eerie in the same way as the howl of a lonely wolf is left to readers to answer. Perhaps each is an expression of a painful longing – to belong? to mate? to be recognized? – but whatever the reason, it is not one that the system feels compelled to share, or even to formulate.

The Google n-grams is also a rich source of pseudo-schemas – copula statements that suggest the equivalence of taxonomically remote ideas – such as "*research* is the *fruit*" (*freq* = 48). The following pair of successive tweets from *@MetaphorMagnet* shows how it can elaborate these "found" objects:

> Remember when research was conducted by prestigious philosophers?
> #Research=#Fruit
>
> Now, research is a fruit eaten only by lowly insects. #Philosopher=#Insect

This metaphor employs schematic reasoning in the mold of Lakoff & Johnson (1980) rather than the analogical mode championed by Gentner *et al*. (1989). The system's propositional knowledge informs it that *philosophers* (among others) conduct *research*, while *insects* (among others) eat *fruit*. This pair of propositions is favored over many other candidates because of the antonymy between *prestigious* (a stereotypical property of *philosophers*) and *lowly* (a stereotypical property of *insects*). Crucially then, and especially so for what follows next, the resulting metaphor rests on an incongruity that is made appropriate in a rhetorical setting that serves to yoke two opposing perspectives on its topic, *research*.

3 All the news that's fit to fingerprint: A Vector Space Approach

Systems such as *@MetaphorMagnet* generate their outputs in a vacuum, without regard for the context in which their metaphors will later be consumed by readers. There are sound reasons for treating a key generator of content as a black-box; for one, *@MetaphorMagnet* is a 3rd-party system of many moving parts that does not invite the low-level tinkering needed to make it context-sensitive to the news; for another, we want our contextualizing matcher to work with potentially many generators of metaphors in a mostly effortless plug-and-play fashion. Our focus then is on competence rather than perfomance. We do not view a metaphor generator as the sum of its procedural mechanisms, but as the sum total of the metaphors it is capable of delivering to the *@MetaphorMirror* system. We thus view the matching of contextless metaphors to contextual headlines as a cross-space mapping problem in which elements of one space, the space of headlines, are mapped to apt elements of another, the space of metaphors.

Let us first consider the headline space. News services cater for a diverse readership by segmenting their offerings along thematic lines. The standard topics – *Sport*, *Politics*, *Business*, *Culture*, etc. – are broad umbrella terms under which a great many stories can shelter. Such coarse-grained classifications serve a useful role in the organization of print newspapers and their online incarnations, but they lack sufficient granularity to support a nuanced mapping of arbitrary metaphors to arbitrary headlines. We can, however, use a topic model, such as *Latent Dirichlet Allocation* (Blei *et al*. 2003), to derive a large, fine-grained set of topics from a news corpus that better reflects the intuitive understanding of events that readers bring with them to a news story. LDA views topics as probabilistic rather than discrete, and generative rather than post-hoc. In constructing a fixed set of tacit and unnamed topics to explain a particualr document set – the precise number of topics is specified by the developer – LDA aims to find the best statistical explanation for the observed lexical similarities between texts in the document set. As any given text will exhibit degrees of affinity to *n* different topics, each topic constitutes a dimension in a vector space in which any text can be represented as an *n*-dimensional vector with a value for each of all n topics. For our news corpus we choose to build an LDA vector space of *n*=100 topics. This corpus contains the full text and headlines of 380 thousand news stories from the Web, gathered between 2000 and 2012 from *Bloomberg* (6%), *Economist* (2%), the *Guardian* (12%), the *Huffington Post* (3%), the *Independent* (7%), the *Irish Times* (7%), the *New York Times* (12%), the *Telegraph* (9%), the *Washington Post*

(8%), *Reuters* (4%) and *Yahoo News* (30%). To this corpus are added 210,000 news tweets from various sources on Twitter, including *@CNNbrk*, *@FOXnews*, *@WSJ* and *@nytimesworld*, harvested between July 2015 and June 2016. When building the LDA model (using the *gensim* implementation of Řehůřek and Sojka, 2010), we used the concatenation of both the lemmatized forms and their POS tags as features of the words in each document for the model.

Let us now consider our metaphor space. As noted above, we employ an extensional rather than an intensional model of this space, and construct it much like our news space: as a large collection of metaphorical micro-texts. These figurative texts are provided by the creators of *@MetaphorMagnet*, who offer large quantities of machine-generated metaphors to other developers for research purposes. This corpus of 22,846,672 metaphors constitutes a wide-ranging sample of *@MetaphorMagnet*'s rhetorical mechanisms and linguistic framings. Once again we use a topic model to capture the tacit themes that recur across these metaphors. However, we do not build a separate topic model for metaphors, but instead build a joint vector space by merging our metaphor corpus with our news corpus. The goal is to accommodate news headlines and metaphors within the very same vector space using the very same topics as dimensions, so that a vector for any given news headline can be directly compared – using a cosine similarity measure – to the vector for any given metaphor. Our choice of the number pf topics is motivated by a desire to achieve an acceptable granularity of themes without encouraging the creation of splinter topics that serve to organize one or another type of content – news or metaphor – but not both together. The success of the joint space depends on the applicability of *all* dimensions to both kinds of content, so instances of each can be meaningfully and incisively compared.

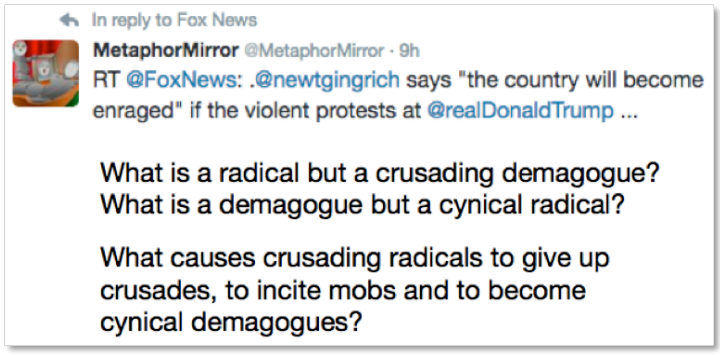
4 *MetaphorMirror* on the wall, what's the most topical trope of all?

We use LDA to construct a common vector space in which metaphors can sit cheek-by-jowl with news headlines, but we could just as easily have used LSA (or *Latent Semantic Analysis*; see Landauer and Dumais, 1997) or Word2Vec (see Mikolov, 2013) to construct our joint space. Our requirements of this space are straightforward: every pre-generated metaphor (all 22,846,672 of them) should be pre-assigned an *n*-dimensional vector of normalized values (where n=100) that reflect a metaphor's affinity to *n* latent topics or themes; these dimensions should be of equal relevance to metaphors and headlines, to foster the clustering of each kind of content in the same pockets of vector space; and each new headline should be quickly mapped to its own *n*-dimensional vector in the space as it arrives, so that it can be compared to all metaphors in the same vicinity of the space using a cosine-similarity measure. This is the matching mechanism at the heart of the *@MetaphorMirror* Twitterbot. As headlines arrive from a range of Twitter sources, such as *@CNNbrk*, *@FOXnews*, *@WSJ*, *@Reuters*, *@nytimesworld*, *@AP* and *@BBCbreaking*, each is reduced to a vector representation (using LDA as our default model) and this vector is compared to that of each metaphor in the space. For efficiency reasons the model *could* insist that each headline must share at least one content word with any metaphor to which it is compared, thus limiting the search for possible comparisons, but the nature of the application is such that an exhaustive search of the metaphor pool is feasible.

@MetaphorMirror aims to tweet one metaphor/headline pairing per hour on average, and by foregoing the need for literal similarity between metaphor and headline the bot can produce content pairings with no lexical overlap at all.

A metaphor is paired with an incoming headline only if the cosine similarity of the two exceeds a minimum threshold, where the default setting of this threshold is 0.9. If no metaphor is found that exhibits this minimum similarity to the headline, no pairing of metaphor to headline is tweeted. There are other conditions under which the system will not produce a pairing. News organizations often tweet the same headline, or near variants thereof, multiple times in the same day; the system will not attempt to pair a metaphor with a headline that has already produced an earlier successful pairing. Likewise, the system strives to avoid repeating itself, and will not select a pairing involving a metaphor that was tweeted by the system in recent memory. If these concerns for repetition mean that no acceptable metaphor can exceed the threshold for an incoming headline, no pairing is produced or tweeted.

The following pairs a metaphor for demagoguery to a headline from *@FOXnews*:

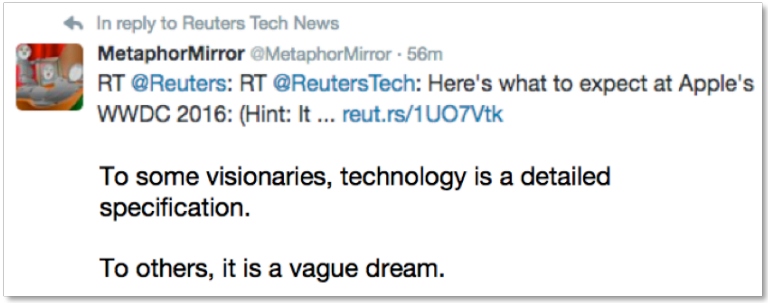


Notice that the metaphor (which is attached as an image rather than a text string so as not to exceed the 140 character limit on tweets) exhibits no lexical overlap with the arriving headline, yet the LDA topic analysis captures the essential similarity between violent political protest and the actions of a cynical demagogue. Because our vector space is distilled from a wide-ranging corpus of news stories in which stereotypical beliefs influence lexical choice, LDA's latent topics allow these implicit stereotypes to guide the selection of metaphors. For example, the bot tweeted the following metaphor on the death of boxer Muhammad Ali, a graceful fighter who became famous for his mantra "dance like a butterfly": dreams, weightless butterflies that you are, how you charm me with your free illusion. Other stereotypes that guide our appreciation of the news, and of a reporter's choice of words and conceits, also exert an influence on the selection of apt metaphors through the dimensions of the LDA model, as in the choice of the following metaphor for the *@WSJ* headline "*IKEA has big ideas for small spaces*":

Dreams, pleasing gifts that you are, how you comfort me with your cheap appeal.

Inevitably, the system cannot but occasionally reflect the biases of its underlying news corpus. For instance, given this *@WSJ* headline following a mass shooting at an Orlando nightclub, "*RT @Sam Walkers Omar Mateen kept threatening to commit mass murder*", the system finds this to be an apt pairing: Guns, roaring monsters that you are, do not menace me with your evil threat. We consider ways of addressing the bias in the underlying news corpora in our concluding remarks.

Stereotypical associations are central to the bot's aim of tweeting apt metaphors:



The association between Apple corp and visionaries with a dream may be the stuff of cliché, but it is on shared beliefs that a compelling metaphor is built. Notice how, in seeking to have it both ways, the metaphor portrays two possible sides of the story: the *positive* dream and the *negative* possibility of disappointment. These metaphors are not chosen to tell readers what to think – to be candid, this is beyond the scope *and* the ability of the system – but simply to encourage them to think more.

5 Empirical Evaluation

We view *@MetaphorMirror* as the linguistic and figurative equivalent of a sommelier that responds to news events with an appropriate pairing of metaphor and headline. Just as a good choice of wine can enrich a fine meal, we designed *@MetaphorMirror* to enrich the user's appreciation of the news with a well-chosen metaphor. Sometimes one trusts the sommelier implicitly, other times one sends the wine back in puzzlement or dismay. So to evaluate the capabilities of *@MetaphorMirror* in its capacity as a proposer of topical metaphors in a changing news environment, we use the crowd-sourcing platform *CrowdFlower* to obtain human judgments on the bot's pairing of metaphors to 90 randomly selected news headlines, plucked from the news in early July 2016. We tested the system using vector spaces built from two corpora – the basic news corpus of full stories and headlines (which spans 2000 to 2012) but no news tweets, and the combination of this full-story corpus with a year of news tweets from Twitter. We refer to the first corpus as *fulltext* and the second as *fulltext +tweets*.

In addition to our topic-model approach we evaluated two other vector-based approaches, based on LSA (Landauer and Dumais, 1997) and Word2Vec (Mikolov, 2013). For the LSA models we again used the *gensim* package of Řehůřek and Sojka (2010) to build 100-dimension compressed vector spaces from each of our two corpora. For the Word2Vec models we used the settings reported in Gatti *et al*. (2015)

for their slogan adaptation system – using the average of the embedding of every word in the sentence as the representation of a whole sentence – also using Google News embeddings for our vocabulary (*https://code.google.com/archive/p/word2vec*). As a baseline we also evaluated wholly random pairings of *@MetaphorMagnet* metaphors to our headlines. It should be noted that this random baseline for metaphor is not the straw man one might expect it to be. Veale (2015) evaluated *@MetaphorMagnet*, by using outputs of another Twitterbot, *@metaphorminute*, as a random baseline. This other bot, from noted Twitterbot creator Darius Kazemi, fills the linguistic template "X is a Y: P and Q" with largely random choices of nouns for X and Y and largely random choices of adjectives for P and Q. When judges on *CrowFlower* were asked to rate the comprehensibility of metaphors from each, Kazemi's bot scored higher than one ought to expect for a random generator, with about 50% of its outputs being rated as moderately to highly comprehensible (compared to 75% for *@MetaphorMagnet*). It seens that framing an idea in the form of a metaphor encourages people to perceive meaning, or the possibility of meaning, where none is actually intended. In the case of our random baseline, we might expect humans who are presented with a metaphor that is grammatically well-formed and semantically coherent to perceive pragmatic resonances with any headline that is paired to it, even if randomly so.

For each pairing of headline and metaphor, and for each condition outlined above, judges were asked to provide ratings for each pairing along three dimensions: *comprehensibility* (of the metaphor and the headline together); *aptness* (of the metaphor to the headline); and *influence* (a self-report of the metaphor's influence on the reader's interpretation and appreciation of the headline). Judges were asked to select values for each dimension from a 5-point (1 to 5) Likert scale. 10 ratings were elicited for each dimension of each pairing under each condition, where the same 90 headlines were paired in each test condition. Table 2 presents the mean values of each dimension under each test condition.

Table 1: *Mean values (+ std. dev.) of each dimension under each pairing condition*

Pairing model	***Aptness***	***Comprehensibility***	***Influence***
LDA (*fulltext+tweets*)	2.95 ± 1.27	3.59 ± 1.05	3.01 ± 1.24
LDA (*fulltext* only)	2.78 ± 1.04	3.54 ± 0.92	2.75 ± 1.03
LSA (*fulltext+tweets*)	2.62 ± 1.10	2.97 ± 1.09	2.44 ± 1.01
LSA (*fulltext* only)	2.40 ± 1.12	2.99 ± 1.15	2.49 ± 1.14
Word2Vec	2.65 ± 0.99	3.38 ± 1.02	2.73 ± 1.00
Random baseline	2.20 ± 1.20	2.54 ± 1.12	2.09 ± 1.24

The LDA model derived from a news corpus of *fulltext* stories (to arm the system with stereotypical associations) and a year of recent news tweets (to condition it to recent world events) outperforms the various other settings of the system across all three dimensions. Though Word2Vec shows a slight improvement over both LSA models, the increase in aptness is not very significant. Table 2 provides significance values for the differences of aptness in Table 1, as calculated using a single-sided Welch t-test (only values for signifcant differences are shown).

Table 2: *Powers of single-sided Welch t-tests between mean aptness ratings for each test condition. Italic numbers show that the power is less than the threshold α=0.05.* ***Bold*** *numbers show that the power is less than the threshold α=0.0001*

	LDA *fulltext only*	**Word2Vec** *Google News*	**LSA** *fulltext +tweets*	**LSA** *fulltext only*	**Random** *baseline no corpus*
LDA *fulltext+tweets*	1.8×10^{-2}	3×10^{-4}	$\mathbf{1.1\times10^{-5}}$	$\mathbf{4\times10^{-7}}$	$\mathbf{1\times10^{-15}}$
LDA *fulltext* only		4.1×10^{-2}	1.8×10^{-2}	$\mathbf{1.6\times10^{-5}}$	$\mathbf{3\times10^{-13}}$
Word2Vec			0.37	8.0×10^{-3}	$\mathbf{3\times10^{-13}}$
LSA *fulltext+tweets*				0.02	$\mathbf{1\times10^{-7}}$
LSA *fulltext only*					$\mathbf{1\times10^{-5}}$

To better understand the distribution of ratings, we placed mean judgments of aptness under all test conditions into four equally-sized bins labeled *Low*, *Average*, *Good* and *Very Good*. Table 3 presents the percentage of headline/metaphor pairings that fall into each bin for human judgments of aptness:

Table 3: *Distribution of mean aptness across 4 quality bins for all test conditions*

Pairing model	***Low***	***Average***	***Good***	***Very Good***
LDA (*fulltext+tweets*)	1.1%	47.8%	**41.1%**	**10%**
LDA (*fulltext* only)	3.3%	**65.6%**	30%	1.1%
LSA (*fulltext+tweets*)	10%	60%	30%	0%
LSA (*fulltext* only)	17.8%	64.4%	16.7%	1.1%
Word2Vec	10%	57.8%	32.2%	0%
Random baseline	**45.5%**	46.7%	6.7%	1.1%

Again the LDA model, derived from a combination of full text stories and recent news tweets, shows the best distribution of results, suggesting that this serve as our default platform in seeking further increases in aptness. In our concluding remarks we next discuss ways in which LDA can be used in a splintered fashion to address issues of bias in the underlying news corpora on which the model is built.

6 Concluding Thoughts and Future Work

Much research has been conducted on the automated analysis of human personality as expressed through one's lexical choices. Chung and Pennebaker (2008), for example, describe an approach and a resource, named the LIWC (*Linguistic Inquery and Word*

*Coun*t) for estimating authorly qualities such as anger, depression, anxiety, affability, positivity, arrogance, analyticality, awareness, topicality (or in-the-moment thinking) and social engagement from a person's textual outputs. An online incarnation of this system (at *www.analyzewords.com*) infers values for these dimensions from the recent tweets of any Twitter account one cares to provide as input. For instance, this tool informs us that *@Oprah* is very upbeat as a Twitter user, while *@realDonaldTrump* is both very upbeat and very angry. Though Twitterbots such as *@MetaphorMagnet* do not pretend to be human – in fact, a key part of their charm for human followers is their combination of overt artificiality *and* meaningfulness to humans – they are nonetheless designed to create and tweet human-quality outputs, and so it is useful to explore what kind of authorly personality they present to the world. The LIWC tool holds no surprises for followers of *@MetaphorMagnet,* however: its personality, based on a sampling of 1,000 words, is deemed to be very angry, very worried, very analytical and very arrogant. While these are interesting qualities for an author to present to readers, the Twitterbot also rates very poorly on the dimension of in-the-moment thinking. So while *@MetaphorMagnet* is capable of generating high-quality metaphors (see e.g. the empirical analysis in Veale, 2015) it is also quite incapable of choosing the best time to use them.

The work and the system described in this paper has sought to address this concern about automated metaphor generation in a way that maintains the modular integrity of the generative component. When developers control both the generator and the contextual adapter, as in the system of Gatti *et al*. (2015) which adapts linguistic expressions to recent news content to generate topical slogans, one is free to make the generative component as context-sensitive as the end-application demands. Nonetheless, the ability to generate large portfolios of creative linguistic artifacts in a vacuum, for contextual reuse at a later time, greatly simplies the issue of topical aptness, allowing developers to focus on competence over performance and to plug and play alternate or additional generative components as desired.

@MetaphorMirror is an attempt to set automated metaphor generation to a topical metronome. Our initial experimental results are sufficiently encouraging to explore more ambitious ways of using metaphor to influence our understanding of topical events in a biased news environment. In this regard we are excited by the possibilities afforded by parallel vector-space modeling. Though we have here used a monolithic vector space distilled from news harvested from a broad spectrum of news providers, it is practical to build individual vector spaces that marry our stock of pre-generated metaphors to news garnered from providers on different ends of the political spectrum. For instance, it is practicable to build right-leaning and left-leaning vector spaces, and to use these spaces to suggest metaphors for headlines originating from providers of opposing views. In this way, *@MetaphorMirror* can pair news headlines from *CNN,* the *BBC* and the *New York Times* with metaphors suggested by a space of *FOXnews* content, or to pair headlines from *FOXnews* with metaphors suggested by a space of content from *CNN,* the *BBC* and the *New York Times*. As Pollio (1997) argues, we often design our metaphors to appear seamless, so as to paper over the rift between competing points of view. Yet there are times that demand metaphors which alert us to the scale of the rift and to the dangers of ignoring it. These times become more numerous and more demanding as our news media becomes more biased.

References

Attardo, S, (2001). Irony as relevant inappropriateness. *Journal of Pragmatics*, **32**:793-826.

Blei, D. M., Ng, A. Y. and Jordan, M. I.. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022.

Brants, T. and Franz, A. (2006). *Web 1T 5-gram v.1.* Linguistic Data Consortium.

Carbonell, J. G. (1981). Metaphor: An inescapable phenomenon in natural language comprehension. *Report 2404*. Pittsburgh, PA: Carnegie Mellon Computer Science Dept.

Chung, C.K. and Pennebaker, J.W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, 46, 96-132.

Fass, D. (1991). Met*: a method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49-90.

Gatti, L., Özbal, G., Guerini, M., Stock, O. and Strapparava, C. (2015). Slogans Are Not Forever: Adapting Linguistic Expressions to the News. In Proc. of *IJCAI'15*, the *24th International Conference on Artificial Intelligence,* pp 2452–2458, Buenos Aires, Argentina. AAAI Press.

Gentner, D., Falkenhainer, B. and Skorstad, J. (1989). Metaphor: The Good, The Bad and the Ugly. In *Theoretical Issues in Natural Language Processing*, Yorick Wilks (Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glucksberg, S. (1998). Understanding metaphors. *Current Directions in Psychological Science,* 7:39-43.

Harmon, S. (2015). FIGURE8: A Novel System for Generating and Evaluating Figurative Language. In *Proc. of the 6th International Conference on Computational Creativity*, Park City, Utah, June 2016.

Hobbs, J. (1981). Metaphor interpretation as selective inferencing. In *Proc. of the 7th International Joint Conference on Artificial Intelligence* - Volume 1, IJCAI'81, pp 85–91, Vancouver, BC, Canada.

Koestler, A. (1964). *The Act of Creation.* Penguin Books.

Lakoff, G. and Johnson, M. (1980). Metaphors We Live By. Illinois: Chicago University Press.

Lakoff, G. (1994) The Master Metaphor List. *http://cogsci.berkeley.edu/*, Uni. of California, Berkeley.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**(2):211-240.

Mason, Z. J. (2004). CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System, *Computational Linguistics*, 30(1):23-44.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv*:1301.3781 [cs], January.

Oring, E. (2003). *Engaging Humor*. University of Illinois Press.

Pollio, H.R. (1996). Boundaries in humor and metaphor. In: Mio, Jeffery Scott and Katz, Albert N. (eds.) *Metaphor, Implications and Applications*, pp 231–253. Mahwah: Lawrence Erlbaum Associates.

Řehůřek, R., and Sojka, P. (2010). Software Framework for Topic Modeling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp 45–50.

Reiter, E. and Dale, R. (2006). Building Natural Language Generation Systems. *Studies in Natural Language Processing*. Cambridge University Press.

Shutova, E. (2010). Metaphor Identification Using Verb and Noun Clustering. In *the Proc. of the 23rd International Conference on Computational Linguistics*, 1001-1010.

Veale, T. and O'Donoghue, D. (2000). Computation and blending. Cognitive Linguistics, 11(3-4): 253–281.

Veale, T. and Hao, Y. (2007). Comprehending and Generating Apt Metaphors: A Web-driven, Case-based Approach to Figurative Language. *In Proc. of AAAI 2007, the 22nd AAAI Conference on Artificial Intelligence.* Vancouver, Canada.

Veale, T., and Li, G. (2011). Creative Introspection and Knowledge Acquisition. In *Proc. of the 25th AAAI Conference on Artificial Intelligence.* San Francisco, CA: AAAI Press.

Veale, T. and Li, G. (2013). Creating Similarity: Lateral Thinking for Vertical Similarity Judgments. *In Proc. of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria*.

Veale, T. (2015). Game of Tropes: Exploring the Placebo Effect in Computational Creativity. In *Proc. of ICCC-2015, the 6th International Conference on Computational Creativity.* Park City, Utah, USA.

Veale, T., Shutova, E. and Beigman Klebanov, B. (2016). *Metaphor: A Computational Perspective.* Morgan Claypool, Synthesis Lectures on Human Language Technologies.

Way, E. C. (1991). Knowledge Representation and Metaphor. *Studies in Cognitive systems*. Holland: Kluwer.

Wilks, Y. (1978). Making Preferences More Active, *Artificial Intelligence* 11.