APPENDIX A
EXPERIMENTAL INSTRUCTIONS: KEYWORD RATING

A. Instructions

In this experiment you will be presented with a news image, an article associated with the image, and a set of keywords describing the image. Your task is to judge how well each of the keywords describe the content of the image given the accompanying article. Some keywords will seem appropriate to you, but others will not. You will make your judgment by choosing a rating from 1 (the keywords are not appropriate) to 7 (the words are appropriate). All keywords were generated automatically by a computer program.

For example, if you were presented with the following image, keywords, and document, you would probably give the keywords *plane* and *airport* a high rating (e.g., 6 or 7) since they are relevant both for the image and the document. Indeed, the image shows a plane in an airport and the article discusses US planes landing in UK airports. Keywords *arms*, *carrying*, and *refuel* should also receive a relatively high rating (e.g., 4-6). Even though they are not explicitly depicted in the image, they are related to the accompanying article which discusses how the British government allowed US planes to refuel in the UK while carrying bombs. Keywords *Scotland* and *book* would probably receive a low rating (e.g., 1 or 2), since they are neither shown in the image nor mentioned in the article.

airport arms carrying plane refuel Scotland book

The British government could face claims it violated international humanitarian laws by allowing US arms flights to Israel to use UK airports.

The Islamic Human Rights Commission (IHRC) is seeking permission to contest government bodies over what it says are crimes against the Geneva Convention. A number of US planes said to be carrying bombs to Israel refueled in the UK during the Lebanon conflict. The IHRC said it received complaints from Britons with families in Lebanon.

The commission is accusing the government of ”grave and serious violations” of international humanitarian law. It is seeking permission to bring its case against the Civil Aviation Authority, the Foreign and Commonwealth Office and Defence Secretary Des Brown in the High Court. The IHRC said it is bringing the case after receiving ”many complaints ... from British citizens whose family members are in Lebanon and facing grave danger as well as acts of terror”. The US aircraft believed to have refueled in the UK are said to have been carrying supplies including ”bunker buster bombs”.

B. Interface

You will be presented with the image, the description keywords, and the document. Once you look at the picture and you read both the document and the keywords, please make your judgment for each keyword by selecting a number between 1 and 7. Each number will be represented by a button, all you have to do is click the button corresponding to your judgment.

There are no ‘correct’ answers, so whatever numbers seem appropriate to you are a valid response. While you are deciding a number for a set of keywords, try to ask the following questions:

- Does the keyword describe information present in the image and the document?
- Does the keyword represent the main topic of the document?
- Does the keyword depict an object present in the picture?

Use high numbers if the answer to the above questions is ‘yes’, low numbers if it is ‘no’, and intermediate numbers for keywords that represent peripheral aspects of the image and document. Try to make up your mind quickly, base your judgments on your first impressions. The experiment will last less than 10 minutes.

APPENDIX B
EXPERIMENTAL INSTRUCTIONS: CAPTION GENERATION

A. Instructions

In this experiment you will be presented with a news image, an article associated with the image, and a caption describing the image. Your task is to judge how well the caption describes the content of the image given the accompanying article and how grammatical the caption is. Some captions will seem appropriate to you, but others will not. You will make your judgement by choosing a rating from 1 (the caption is not appropriate) to 7 (the caption is appropriate). All captions were generated automatically by a computer program.

For example, if you were presented with the following document, image, and caption:

A US plane landed in a UK airport with bunker buster bombs.

The British government could face claims it violated international humanitarian laws by allowing US arms flights to Israel to use UK airports.

The Islamic Human Rights Commission (IHRC) is seeking permission to contest government bodies over what it says are crimes against the Geneva Convention. A number of US planes said to be carrying bombs to Israel refueled in the UK during the Lebanon conflict. The IHRC said it received complaints from Britons with families in Lebanon.

The commission is accusing the government of ”grave and serious violations” of international humanitarian law. It is seeking permission to bring its case against the Civil Aviation Authority, the Foreign and Commonwealth Office and Defence Secretary Des Brown in the High Court. The IHRC said it is bringing the case after receiving ”many complaints ... from British citizens whose family members are in Lebanon and facing grave danger as well as acts of terror”. The US aircraft believed to have refueled in the UK are said to have been carrying supplies including ”bunker buster bombs”.

You would probably give the caption in bold a higher content rating (e.g., 6 or 7) since it is relevant both for the image and the document. Indeed, the image shows a plane in an airport and the article discusses US planes landing in UK airports with bombs. Even though the words bombs, US and UK are not explicitly depicted in the image, they are related to the accompanying article which discusses how the British government allowed US planes to refuel in the UK while carrying

bombs to Israel. If a caption is neither related to the image nor to the article, then it should receive a lower content rating. If the caption is grammatical, then you should rate it a higher content score. If it is not fluent and reads like word salad, then you should give it a lower rate.

B. Interface

You will be presented with the document, the image, and the caption. Once you read the document, look at the picture and read its caption, please make your judgement by selecting a number between 1 and 7. Each number will be represented by a button, all you have to do is click the button corresponding to your judgement.

There are no 'correct' answers, so whatever numbers seem appropriate to you are a valid response. While you are deciding the rating, try to ask the following questions:

- Does the caption describe information present in the image and the document?
- Does the caption represent the main topic of the document?
- Does the caption depict an object present in the picture?
- Does the caption seem fluent? Is it understandable?

Use high numbers if the answer to the above questions is 'yes', low numbers if it is 'no', and intermediate numbers for captions that represent peripheral aspects of the image and document. Try to make up your mind quickly, base your judgments on your first impressions. The experiment will last less than 10 minutes.

APPENDIX C
EXAMPLES OF SYSTEM OUTPUT

Table I provides examples of system output (on the test set). Specifically, we show the image keywords provided by TxtLDA and the corresponding caption generated by the abstractive phrase-based model (A_P). For comparison, we also include the original human-authored caption. We omit the document for the sake of brevity.

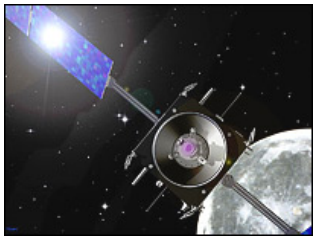
As can be seen, in most cases the caption mentions an object depicted in the image. For instance, caption (a) mentions the moon and the smart 1 probe, caption (b) mentions Daniel McGurk, caption (c) mentions children eating, caption (d) water, caption (f) the Chelsea player William Gallas, caption (d) a mobile phone, caption (h) a group of immigrants and so on. The identification of these objects is aided by taking visual features into account, e.g., there are several images of mobile phones in our database (see caption (g)), children (see captions (i) and (c)), men (see caption (b)) women (see caption (i)), and Tony Blair (see caption (e)). Less frequently occurring objects (e.g., the smart probe 1), persons (e.g., Danny McGurk or William Gallas) or actions (e.g., eat, discuss, breastfeed) are highlighted by analyzing the document and by being able to infer what it is about.

In most cases the automatically generated captions are grammatical. Examples of relatively incoherent output are captions (c), (f), and (i). This is due to the fact that the phrase-based model does not have a notion of global grammatical coherence. Consider caption (c) for example. Here, the individual phrases *the children of*, *good Ayrshire produce*, *enjoy eating*, and *shows in their school dinners* are well-formed but their combination yields an awkward sentence. The model has no notion of selectional restrictions, i.e., it does not know that *produce* cannot have *children*; it also does not take number agreement into account, and thus associates words in plural form (e.g., *the children*) with singular words (e.g., *shows*).

It is interesting to note that the generation model creates captions based on the keywords suggested by MixLDA (keywords figuring in the caption are shown with boldface in Table I). This indicates that the captions generated by our model are *grounded* in textual and visual information. Even though we consider the output of MixLDA indirectly as a form of smoothing for the language model performing surface realization, we observe that it biases generation towards the selection of words deemed important by content selection. Finally, note that the generated caption is often different from the human-authored gold standard. In some cases the two captions express similar content, albeit using different words (see the captions for images (a), (b), (d), and (f)). The discrepancy between human-written and automatic captions is not surprising given the creative nature of the caption creation task and the fact that the relationship between image and caption is relatively loose. Words that are shared between the automatic and original caption are underlined in Table I (we only indicate overlap amongst content words).

TABLE I

KEYWORDS SUGGESTED BY OUR MIXLDA MODEL AND CAPTIONS WRITTEN BY HUMANS (G) AND GENERATED BY OUR PHRASE-BASED ABSTRACTIVE MODEL (A_P). BOLDFACE WORDS ARE SHARED BETWEEN MIXLDA AND A_P. UNDERLINED WORDS ARE SHARED BETWEEN A_P AND G.

	a.	b.	c.
			
MixLDA	moon, probe, end, surface, space, map, mission, professor, impact, crash	man, death, murder, tell, home, court, investigation, Belfast, area, result	school, local, Ayrshire, quality, eat, child, council, pilot, meal, minister
A_P	The smart 1 **probe** produced detailed **maps** of the **moon**.	The **court** heard Danny McGurk was **murdered** in the **area**.	The **children** of good **Ayrshire** produce enjoying **eating** shows in their **school** dinners.
G	Scientists hope to view the rock beneath the moon's surface.	Daniel McGurk was murdered in August 2003.	Schools are being urged to source food locally.

	d.	e.	f.
			
MixLDA	water, customer, firm, England, complaint, service, system, rise, increase, council	meeting, terrorism, India, minister, man, London, Blair, discuss, bombing, European	play, Chelsea, want, club, season, new, player, goal, score, manager
A_P	The consumer **council** for **water** made a rise in **complaints** across the UK.	The two **men** discussed the fight against **terrorism** with **European** union leaders.	The problems for **Chelsea** William Gallas made towards the end of last **season**.
G	Complaints about the water quality from taps fell.	Mr Blair visited India last September.	Gallas failed to meet up with Chelsea on their pre-season tour.

	g.	h.	i.
			
MixLDA	mobile, network, funding, block, charter, crime, street, industry, steal, phone	immigrant, island, group, illegal, sea, captain, die, boat, rescue, report	child, mother, baby, study, woman, health, breastfeed, researcher, intelligent, family
A_P	The majority of **mobile phones stolen** in the **charter**.	The **group** of **illegal immigrants rescued** by the Italian navy.	A **study** found **breastfeeding mothers breastfed children** showed.
G	It is hoped that the pledge will put people off buying stolen phones.	Many immigrants die or drown trying to cross to Europe by sea every year.	Breastfed babies tend to be brighter.

	j.	k.	l.
			
MixLDA	Diana, police, case, crash, princess, report, death, inquest, Paris, Burgess	gun, old, shoot, video, boy, rally, song, live, tv, father	car, incident, hear, pavement, place, dead, tell, old, drive, dispute
A_P	The findings into the **deaths** of **Diana**, **princess** of Wales.	The **father** of a Scottish toddler talks about the devastation caused by **guns**.	The **car** at the bottom of the road involved in a **dispute**.
G	Princess Diana died in a car crash in Paris in 1997.	David and Ozlem Grimason's son was killed by a stray bullet.	The car struck the child and his mother after mounting the pavement.