Finding Similar Content Within Different Documents

**John O. Everett, Daniel G. Bobrow, Cleo Condoravdi,
Richard Crouch, Valeria de Paiva, Reinhard Stolle**
PARC
3333 Coyote Hill Road
Palo Alto, California 94304
{jeverett, bobrow, condorav, crouch, paiva, rstolle}@parc.com

Documents about a particular subject may describe the same phenomena in different words. For example, in a database of tips for repairing photocopiers, there are two tips about a safety cable failure. One describes the situation as "the cable is too stiff, which causes it to snap. Remove the sleeve from the cable." The other says "Stripping the cover from the cable makes it more flexible."

Assessing the similarity of texts requires the application of knowledge about language and about the world. In this case, we need to know, for example, that "stiff" and "flexible" both refer to the rigidity of an object, and that one is the inverse of the other, and also that a sleeve is a type of covering. Our research focuses on combining deep natural language analysis with domain knowledge representations.

We are developing a layered approach to automatically identifying similar document content. The first layer extracts from the text semantically normalized entities (in our case things like parts, e.g., *photoreceptor belt*) and relevant activities (in our case higher level concept representing domain specific actions, such as *cleaning*). The set of normalized entities can be used as a signature for identifying tips likely to contain information about the same topic.

The second layer builds from these normalized entities and activities representation fragments that correspond to events, which are actions applied to specific entities (e.g., *cleaning* a *photoreceptor belt*). These can be used to identify parts of a pair of related tips that have similar content. Matches between representations at this level are reliable and precise, as the representations are independent of the particular words of the text.

The third layer links these events with causal or temporal relations, to approximate the macro structure of the text. The partial matches and the higher level structure enable the generation of hypotheses about the relation of other parts of the tips not previously matched, such as identifying an action sequence as a workaround for repairing a machine, and placing it in correspondence with an instruction sequence for a standard method for fixing the machine.

This layered approach allows us to build up increasingly refined relationships among similar documents, providing some information about similarity of tips at each level. As a result, performance degrades gracefully in the face of ambiguous or incomplete natural language analyses.